Bounding the generalization error of convex combinations of classifiers: balancing the dimensionality and the margins

Vladimir Koltchinskii [*], Dmitriy Panchenko
Department of Mathematics and Statistics
The University of New Mexico
and Fernando Lozano
Department of Electrical and Computer Engineering
The University of New Mexico

October 19, 2000

Abstract

A problem of bounding the generalization error of a classifier $f \in \mathrm{conv}(\mathcal{H})$, where $\mathcal{H}$ is a "base" class of functions (classifiers), is considered. This problem frequently occurs in computer learning, where efficient algorithms of combining simple classifiers into a complex one (such as boosting and bagging) have attracted a lot of attention. Using Talagrand's concentration inequalities for empirical processes, we obtain new sharper bounds on the generalization error of combined classifiers that take into account both the empirical distribution of "classification margins" and an "approximate dimension" of the classifiers and study the performance of these bounds in several experiments with learning algorithms.

1991 AMS subject classification: primary 62G05, secondary 62G20, 60F15
Keywords and phrases: generalization error, combined classifier, margin, approximate dimension, empirical process, Rademacher process, random entropies, concentration inequalities, boosting, bagging
Short title: Dimensionality and Margins

[*]Partially supported by NSA Grant MDA904-99-1-0031

1 Introduction

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a sample of n labeled training examples that are independent identically distributed copies of a random couple (X, Y), X being an "instance" in a measurable space S and Y being a "label" taking values in $\{-1, 1\}$. Let P denote the distribution of the couple (X, Y). Given a measurable function f from S into $\mathbb{R}$, we use $\mathrm{sign}(f(x))$ as a predictor of the unknown label of an instance $x \in S$. We will call f a classifier of the examples from S. The quantity $\mathbb{P}\{Yf(X) \leq 0\} = P\{(x, y) : yf(x) \leq 0\}$ is called *the generalization error* of the classifier f. The goal of learning (classification) is, given a set of training examples, to find a classifier f with a small generalization error.

Some of the important recent advances in statistical learning theory are related to the development of complex classifiers that are combinations of simpler ones. In so called *voting methods* of combining classifiers (such as boosting, bagging, etc.) a complex classifier produced by a learning algorithm is a convex combination of simpler classifiers from the base class.

Let $\mathcal{H}$ be a class of functions from S into $\mathbb{R}$ (base classifiers) and let $\mathcal{F} := \mathrm{conv}(\mathcal{H})$ denote the symmetric convex hull of $\mathcal{H}$:

$$\mathrm{conv}(\mathcal{H}) := \Big\{\sum_{i=1}^{N} \lambda_i h_i : N \geq 1, \lambda_i \in \mathbb{R}, \sum_{i=1}^{N} |\lambda_i| \leq 1,\ h_i \in \mathcal{H}\Big\}.$$

Our main goal in this paper is to develop new probabilistic upper bounds on the generalization error of a classifier f from the symmetric convex hull $\mathcal{F} = \mathrm{conv}(\mathcal{H})$ of the base class. The well known approach to such a problem, developed in pathbreaking works of Vapnik and Chervonenkis (see (Vapnik, 1998) and references therein), is based on an easy bound

$$P\{(x, y) : yf(x) \leq 0\} \leq P_n\{(x, y) : yf(x) \leq 0\} + \sup_{C \in \mathcal{C}} [P(C) - P_n(C)],$$

where P_n is the empirical distribution of the training examples, i.e. for any set $C \subset S \times \{-1, 1\}$, $P_n(C)$ is the frequency of training examples in the set C,

$$\mathcal{C} := \Big\{\{(x, y) : yf(x) \leq 0\} : f \in \mathcal{F}\Big\},$$

and on further bounding the uniform (over the class $\mathcal{C}$) deviation of the empirical distribution P_n from the true distribution P. The methods that are used to solve this problem belong to the theory of empirical processes and the crucial role is played by the VC-dimension of the class $\mathcal{C}$, or by more sophisticated entropy characteristics of the class. For instance, if $m^{\mathcal{C}}(n)$ denotes the maximal number of subsets obtainable by intersecting a sample of size n with the class $\mathcal{C}$ (the so called shattering number), then the following bound holds (see (Devroye et al., 1996), Theorem 12.6) for all $\varepsilon > 0$

$$\mathbb{P}\Big\{P\{(x, y) : yf(x) \leq 0\} \geq P_n\{(x, y) : yf(x) \leq 0\} + \varepsilon\Big\} \leq 8m^{\mathcal{C}}(n)e^{-n\varepsilon^2/32}.$$

It follows from this bound that the training error measures the generalization error of a classifier $f \in \mathcal{F}$ with the accuracy $O\Big(\sqrt{\frac{V(\mathcal{C})\log n}{n}}\Big)$, where $V(\mathcal{C})$ is the VC-dimension of the

class $\mathcal{C}$. In the so called zero-error case, when there exists a classifier $\hat{f} \in \mathcal{F}$ with zero training error, we even have the bound (see (Devroye et al., 1996), Theorem 12.7):

$$\mathbb{P}\Big\{P\{(x,y) : y\hat{f}(x) \le 0\} \ge \varepsilon\Big\} \le 2m^{\mathcal{C}}(2n)2^{-n\varepsilon/2},$$

which implies that the generalization error of the classifier $\hat{f}$ is of the order $O\Big(\frac{V(\mathcal{C})\log n}{n}\Big)$. The above bounds, however, do not apply directly to the case of the class $\mathcal{F} = \mathrm{conv}(\mathcal{H})$, which is of interest in applications to bounding the generalization error of the voting methods, since in this case typically $V(\mathcal{C}) = +\infty$. Even when one deals with a finite number of base classifiers in a convex combination (which is the case, say, with boosting after finite number of rounds), the VC-dimensions of the classes involved are becoming rather large, so the above bounds do not explain the generalization ability of boosting and other voting methods observed in numerous experiments. This motivated Bartlett (Bartlett, 1998), Schapire, Freund, Bartlett and Lee (Schapire et al., 1998) (see also (Anthony and Bartlett, 1999)) to develop a new class of upper bounds on generalization error of a convex combination of classifiers, expressed in terms of empirical distribution of margins (the role of classification margins in improving the generalization ability of learning machines was clear in earlier work on support vector machines as well, see (Cortes and Vapnik, 1995). The margin of a classifier f on a training example (X,Y) is defined as the product $Yf(X)$. Schapire, Freund, Bartlett and Lee (Schapire et al., 1998) showed that for a given $\alpha \in (0,1)$ with probability at least $1-\alpha$ for all $f \in \mathrm{conv}(\mathcal{H})$

$$P\{(x,y) : yf(x) \le 0\} \le \inf_{\delta}\Big[P_n\{(x,y) : yf(x) \le \delta\} + \frac{C}{\sqrt{n}}\Big(\frac{V(\mathcal{H})\log^2(\frac{n}{V(\mathcal{H})})}{\delta^2} + \log(1/\alpha)\Big)^{1/2}\Big].$$

Choosing in the above bound the value of $\delta = \hat{\delta}(f)$ that solves the equation

$$\delta P_n\{(x,y) : yf(x) \le \delta\} = \sqrt{\frac{V(\mathcal{H})}{n}}$$

(which is nearly an optimal choice), one gets (ignoring the logarithmic factors) the generalization error of a classifier f from the convex hull of the order

$$O\Big(\frac{1}{\hat{\delta}(f)}\sqrt{\frac{V(\mathcal{H})}{n}}\Big).$$

Koltchinskii and Panchenko (Koltchinskii and Panchenko, 1999), using the methods of the theory of Empirical, Gaussian and Rademacher Processes (concentration inequalities, symmetrization, comparison inequalities) generalized and refined this type of bounds. They also suggested a way to improve these bounds under certain assumptions on the growth of random entropies of a class $\mathcal{F}$ to which the classifier belongs. The new bounds are based on the notion of γ-margin of the classifier, introduced in their paper. The γ-margins are defined for $\gamma \in (0,1)$ (see the definitions in Section 2 below), the value of $\gamma = 1$ roughly corresponds to the case studied in (Schapire et al., 1998). The quality of the bound improves as γ decreases to 0. However, the bounds of this type are proved to hold for the values of $\gamma \ge 2\alpha/(2+\alpha)$,

where $\alpha \in (0,2)$ is the growth exponent of the random entropy of the class $\mathcal{F}$. In the case of $\mathcal{F} := \mathrm{conv}(\mathcal{H})$, where $\mathcal{H}$ is a VC-class with VC-dimension $V(\mathcal{H})$, this leads to the values of $\alpha = 2(V(\mathcal{H})-1)/V(\mathcal{H}) < 2$, which allows one to use γ-margins with $\gamma < 1$ (but it is going to be rather close to 1 unless the VC-dimension is very small). The experiments of Koltchinskii, Panchenko and Lozano (Koltchinskii et al., 2000) showed that, in the case of the classifiers obtained in consecutive rounds of boosting, the bounds on the generalization error in terms of γ-margins hold even for much smaller values of γ. This allows one to conjecture that such classifiers belong, in fact, to a class $\mathcal{F} \subset \mathrm{conv}(\mathcal{H})$ whose entropy might be much smaller than the entropy of the whole convex hull. The problem, though, is that it is practically impossible to identify such a class prior to experiments, leaving the question of how to choose the values of γ for which the bounds hold open. In this paper, we develop a new approach to this problem. Namely, we suggest an adaptive bound on the generalization error of a convex combination of classifiers from a base class that is based on the one hand on the margins of the combined classifiers and on the other hand on their *approximate dimensions* (the numbers of "large enough" coefficients in the convex combinations). This adaptive bound "captures" the size of the entropy of a subset of the convex hull to which the classifier actually belongs.

The results are formulated precisely in Section 2. The proofs that heavily rely upon Talagrand's concentration and deviation inequalities for empirical processes are given in section 3. Section 4 includes the results of several experiments with existing learning algorithms (such as boosting and bagging) for which we computed the bounds on the learning curves that follow from our results. We also discuss here some approaches to combining classifiers that attempt to minimize the margin cost function keeping the dimension of the classifier small.

2 Empirical margins and approximate dimensions: main results

Let $(S, \mathcal{A})$ be a measurable space and let $\mathcal{F}$ be a class of measurable functions on $(S, \mathcal{A})$. In this section, in order to shorten the notations, we suppress the labeles. If one wants to apply the results in the setting of the Introduction, one has to consider instead of S the space $S \times \{-1,1\}$ and instead of a function f on S, a function $(x,y) \mapsto yf(x)$ on $S \times \{-1,1\}$. The results can be also used in the case of multiclass problems (see Section 5 in (Koltchinskii and Panchenko, 1999)). In what follows P denotes a probability measure on $(S, \mathcal{A})$, $\{X_n\}$ is a sequence of i.i.d. random variables, defined on a probability space $(\Omega, \Sigma, \mathbb{P})$ and taking values in $(S, \mathcal{A})$ with distribution P, P_n denote the empirical measure based on the sample $(X_1, \ldots, X_n)$:

$$P_n(A) := n^{-1} \sum_{i=1}^{n} I_A(X_i), \ A \subset S.$$

We start with extending the bounds on generalization error, obtained by Koltchinskii and Panchenko (Koltchinskii and Panchenko, 1999) in terms of so called γ-margins.

Below we give a definition of what we call $\psi-$bounds that will play a major role in

bounding the generalization error of classifiers. These quantities depend on a function ψ that will characterize the complexity of the class $\mathcal{F}$, and therefore determine the quality of the bounds.

Let ψ be a concave nondecreasing function on $[0, +\infty)$ with $\psi(0) = 0$. For a fixed $\varepsilon > 0$, denote by $\delta_n^\psi(\varepsilon)$ the largest solution of the equation

$$\varepsilon = \frac{1}{\delta\sqrt{n}}\psi(\delta\sqrt{\varepsilon}) \tag{2.1}$$

(if ψ is strictly concave, the solution of the equation (2.1) is unique). Clearly, for a concave ψ the function $\varphi(x) \equiv \frac{\psi(x)}{x}$ is nonincreasing. Therefore, it is easy to see that

$$\delta_n^\psi(\varepsilon) = \frac{\varphi^{-1}(\sqrt{\varepsilon n})}{\sqrt{\varepsilon}}.$$

Given a function f and $t > 0$, define the following quantity

$$\varepsilon_n^\psi(f;t) := \inf\Big\{\varepsilon \geq \frac{t \bigvee 2\log n}{n} : P\{f \leq \delta_n^\psi(\varepsilon)\} \leq \varepsilon\Big\}$$

and its empirical version

$$\hat{\varepsilon}_n^\psi(f;t) := \inf\Big\{\varepsilon \geq \frac{t \bigvee 2\log n}{n} : P_n\{f \leq \delta_n^\psi(\varepsilon)\} \leq \varepsilon\Big\}$$

Since for all $\varepsilon > 0$, $\delta_n^\psi(\varepsilon) \geq 0$, it immediately follows from the definition that for all $f \in \mathcal{F}$

$$P\{f \leq 0\} \leq \inf\{P\{f \leq \delta_n^\psi(\varepsilon)\} : \varepsilon \geq \varepsilon_n^\psi(f;t)\} \leq \varepsilon_n^\psi(f;t).$$

We will call $\varepsilon_n^\psi(f;t)$ and $\hat{\varepsilon}_n^\psi(f;t)$ *the ψ-bound* and *the empirical ψ-bound* of the classifier f, respectively. We show below that under a proper assumption on the random entropy of the class $\mathcal{F}$, with a high probability the empirical ψ-bounds $\hat{\varepsilon}_n^\psi(f;t)$ are, for all the functions from the class, within a multiplicative constant from the true ψ-bounds $\varepsilon_n^\psi(f;t)$. This allows one to replace $\varepsilon_n^\psi(f;t)$ in the above bound on $P\{f \leq 0\}$ by $\hat{\varepsilon}_n^\psi(f;t)$ (which gives in applications a bound on the generalization errors of classifiers).

Given a metric space (T, d), we denote $H_d(T;\varepsilon)$ the ε-entropy of T with respect to d, i.e.

$$H_d(T;\varepsilon) := \log N_d(T;\varepsilon),$$

where $N_d(T;\varepsilon)$ is the minimal number of balls of radius ε covering T. If Q is a probability measure on $(S;\mathcal{A})$, $d_{Q,2}$ will denote the metric of the space $L_2(S;dQ) : d_{Q,2}(f;g) := (Q|f - g|^2)^{1/2}$.

Theorem 1 *Let ψ be a concave nondecreasing function on $[0, +\infty)$ with $\psi(0) = 0$. Suppose the following bound on Dudley's entropy integral holds with some $D_n > 0$:*

$$\int_0^x H^{1/2}_{d_{P_n,2}}(\mathcal{F}, u)du \leq D_n\psi(x), \ \ x > 0 \text{ a.s.} \tag{2.2}$$

where $D_n = D_n(X_1, \dots, X_n)$ is a function of training examples such that $\mathbb{E}D_n < \infty$. Then there exist absolute constants $A, B > 0$ such that for $\bar{A} := A(1 + \mathbb{E}D_n)^2$ and for all $t > 0$

$$\mathbb{P}\Big\{\forall f \in \mathcal{F} : \bar{A}^{-1}\hat{\varepsilon}_n^{\psi}(f;t) \le \varepsilon_n^{\psi}(f;t) \le \bar{A}\hat{\varepsilon}_n^{\psi}(f;t)\Big\}$$
$$\ge 1 - B\log_2\log_2\frac{n}{t\bigvee 2\log n}\exp\Big\{-\big(\frac{t}{2}\bigvee\log n\big)\Big\}. \tag{2.3}$$

The following corollary is immediate.

Corollary 1 *Under the conditions of Theorem 1 there exist numerical constants $A, B > 0$ such that for $\bar{A} := A(1 + \mathbb{E}D_n)^2$ and for all $t > 0$*

$$\mathbb{P}\Big\{\exists f \in \mathcal{F} : P\{f \le 0\} \ge \bar{A}\hat{\varepsilon}_n^{\psi}(f;t)\Big\} \le B\log_2\log_2\frac{n}{t\bigvee 2\log n}\exp\Big\{-\big(\frac{t}{2}\bigvee\log n\big)\Big\}. \tag{2.4}$$

Example 1. Let $\alpha \in (0,2)$ and $\psi(x) \equiv x^{1-\alpha/2}$. Let $\gamma := \frac{2\alpha}{\alpha+2}$. Koltchinskii and Panchenko (Koltchinskii and Panchenko, 1999) defined γ-margins of a function f as follows:

$$\delta_n(\gamma; f) := \sup\Big\{\delta \in (0,1) : \delta^{\gamma}P\{f \le \delta\} \le n^{-1+\frac{\gamma}{2}}\Big\},$$
$$\hat{\delta}_n(\gamma; f) := \sup\Big\{\delta \in (0,1) : \delta^{\gamma}P_n\{f \le \delta\} \le n^{-1+\frac{\gamma}{2}}\Big\}.$$

An easy computation shows that

$$\varepsilon_n^{\psi}(f; n^{\gamma/2}) = \frac{1}{n^{1-\gamma/2}\delta_n(\gamma; f)^{\gamma}}.$$

Corollary 1 immediately implies that if for some $\alpha \in (0,2)$ and $D_n > 0$, $\mathbb{E}D_n < \infty$

$$H_{d_{n,2}}(\mathcal{F}; u) \le D_n^2 u^{-\alpha},\ u > 0 \text{ a.s.},$$

then for any $\gamma \ge \frac{2\alpha}{\alpha+2}$ there exist constants $A, B > 0$ such that for $\bar{A} := A(1 + \mathbb{E}D_n)^2$

$$\mathbb{P}\Big\{\exists f \in \mathcal{F} : P\{f \le 0\} \ge \frac{\bar{A}}{n^{1-\gamma/2}\hat{\delta}_n(\gamma; f)^{\gamma}}\Big\} \le B\log_2\log_2 n\exp\Big\{-n^{\gamma/2}/2\Big\} \tag{2.5}$$

(see also (Koltchinskii and Panchenko, 1999)). It is easy to see that the quantity

$$\frac{1}{n^{1-\gamma/2}\hat{\delta}_n(\gamma; f)^{\gamma}} \tag{2.6}$$

in the above upper bound on the generalization error *becomes smaller* as γ decreases from 1 to 0. The Schapire-Freund-Bartlett-Lee type of bounds correspond to the worst choice of γ ($\gamma = 1$). In the case when $\mathcal{F}$ is the symmetric convex hull of a VC-class $\mathcal{H}$ with VC-dimension $V(\mathcal{H})$ the value of α is equal to $\frac{2(V(\mathcal{H})-1))}{V(\mathcal{H})} < 2$ that allows us to have $\gamma < 1$, improving the previously known bound. In fact, Koltchinskii, Panchenko and Lozano (Koltchinskii et al.,

2000) computed the empirical γ-margins of classifiers obtained in consecutive rounds of boosting and observed that the bounds on their generalization error in terms of γ-margins hold even for much smaller values of γ. This allows one to conjecture that such classifiers belong, in fact, to a class $\mathcal{F} \subset \mathrm{conv}(\mathcal{H})$ whose entropy might be much smaller than the entropy of the whole convex hull.

Example 2. Consider now the case of $\psi(x) \equiv x\sqrt{\log \frac{e}{x}}$ for $x \leq 1$ and $\psi(x) \equiv x$ for $x > 1$. Then, by a simple computation,

$$\delta_n^{\psi}(\varepsilon) = \frac{e^{1-n\varepsilon}}{\sqrt{\varepsilon}},\ \varepsilon \geq n^{-1}.$$

If we define

$$\hat{\varepsilon}_n^{VC}(f;t) := \inf\Big\{\varepsilon \geq \frac{t \bigvee 2\log n}{n} : P_n\{f \leq \frac{e^{1-n\varepsilon}}{\sqrt{\varepsilon}}\} \leq \varepsilon\Big\}, \tag{2.7}$$

then under the condition

$$H_{d_{P_n,2}}(\mathcal{F};u) \leq D_n^2 \log \frac{1}{u} \bigvee 1,\ u > 0 \text{ a.s.},$$

with some $D_n = D_n(X_1,\dots,X_n)$, $\mathbb{E}D_n < +\infty$ (which holds, for instance, if $\mathcal{F}$ is a VC-subgraph class), we get from Corollary 1 that with some numerical constants $A, B > 0$ for all $t > 0$

$$\mathbb{P}\Big\{\exists f \in \mathcal{F} : P\{f \leq 0\} \geq \bar{A}\hat{\varepsilon}_n^{VC}(f;t)\Big\} \leq B \log_2 \log_2 \frac{n}{t \bigvee 2\log n} \exp\Big\{-\big(\frac{t}{2} \bigvee \log n\big)\Big\},$$

where $\bar{A} := A(1 + \mathbb{E}D_n)^2$.

The proofs of Theorem 1 and Theorem 3 below are based on the following generalization of one of the results of Koltchinskii and Panchenko (Koltchinskii and Panchenko, 1999) (that itself relies heavily on the concentration inequality for empirical processes due to Talagrand).

Given a nondecreasing concave function ψ on $[0, +\infty)$ with $\psi(0) = 0$ and a fixed number $\delta > 0$, we denote by $\varepsilon_n^{\psi}(\delta) > 0$ the smallest solution of the equation (2.1) with respect to ε.

Theorem 2 *Suppose that condition (2.2) holds with some concave nondecreasing ψ such that $\psi(0) = 0$. Then, for all $\delta > 0$ and for all $\varepsilon \geq \varepsilon_n^{\psi}(\delta) \vee \frac{2\log n}{n}$ the following bounds hold*

$$\mathbb{P}\Big\{\exists f \in \mathcal{F}\ P_n\{f \leq \delta\} \leq \varepsilon \text{ and } P\{f \leq \frac{\delta}{2}\} \geq \bar{A}\varepsilon\Big\} \leq$$

$$\leq B \log_2 \log_2 \varepsilon^{-1} \exp\{-\frac{n\varepsilon}{2}\}.$$

and

$$\mathbb{P}\Big\{\exists f \in \mathcal{F}\ P\{f \leq \delta\} \leq \varepsilon \text{ and } P_n\{f \leq \frac{\delta}{2}\} \geq \bar{A}\varepsilon\Big\} \leq$$

$$\leq B \log_2 \log_2 \varepsilon^{-1} \exp\{-\frac{n\varepsilon}{2}\},$$

where $\bar{A} = A(1 + \mathbb{E}D_n)^2$ and A, B are numerical constants.

There are two major problems with the margin type bounds, given above. First of all, the values of the constants involved in the bounds are far from being optimal and are too large at the moment. Their improvement is related to a hard problem of optimizing the constants in Talagrand's concentration inequalities for empirical and Rademacher processes, used in the proofs below. However, in the case when $\mathcal{F} = \mathrm{conv}(\mathcal{H})$ the constants in question depend only on the base class $\mathcal{H}$ and this allows one to use the bounds to study the behavior of the generalization error when the the number of rounds of learning algorithms (such as boosting) increases. Another problem is related to the fact that there is no much prior knowledge about the subset of $\mathrm{conv}(\mathcal{H})$ to which a classifier created by boosting or another method of combining the classifiers is going to belong. This makes one to use the value of

$$\gamma = \frac{2\alpha}{\alpha + 2} = \frac{2(V(\mathcal{H}) - 1)}{2V(\mathcal{H}) - 1} \tag{2.8}$$

which is very close to 1 unless the VC-dimension of the base is *very* small. Our major goal in the current paper is to address this problem. We do this by proving a new upper bound on the generalization error of a classifier that belongs to a convex hull of a base class. The bound includes the sum of two main terms. The first one is an "approximate" dimension" of the classifier (the number of "large enough" coefficients in the convex combination) divided by the sample size. The second term is related to the margins of the classifier. Balancing these two terms allows us to get rather tight upper bound that "captures" the size of the entropy of a class to which the classifier actually belongs. It combines previously known bounds in terms of VC-dimension (in zero-error case) and in terms of margins and becomes close to one of these two bounds in the extreme cases.

Let $\mathcal{H}$ be a class of measurable functions from $(S, \mathcal{A})$ into $\mathbb{R}$. Let $\mathcal{F} \subset \mathrm{conv}(\mathcal{H})$. For a function $f \in \mathcal{F}$ and a number $\Delta \in [0, 1]$, we define *the approximate Δ-dimension* of f as the integer number $d \geq 0$ such that there exist $N \geq 1$, functions $h_j \in \mathcal{H},\ j = 1, \dots, N$ and numbers $\lambda_j \in \mathbb{R},\ j = 1, \dots, N$ satisfying the conditions $f = \sum_{j=1}^{N} \lambda_j h_j$, $\sum_{j=1}^{N} |\lambda_j| \leq 1$ and $\sum_{j=d+1}^{N} |\lambda_j| \leq \Delta$. The Δ-dimension of f will be denoted by $d(f; \Delta)$. Note that this definition depends on the representation $f = \sum \lambda_j h_j$, and one is free to use any but the choice that produces smaller $d(f; \Delta)$ is advantageous.

In what follows we assume that for some $V > 0$ and $K > 0$ and for all probability measures Q on $(S; \mathcal{A})$

$$N_{d_{Q,2}}(\mathcal{H}; (QH^2)^{\frac{1}{2}}\varepsilon) \leq K\varepsilon^{-V},\ \varepsilon > 0, \tag{2.9}$$

where H is a measurable envelope of $\mathcal{H}$. In particular, this condition holds if $\mathcal{H}$ is a VC-subgraph class. This condition implies the bound on the entropy

$$H_{d_{Q,2}}(\mathrm{conv}(\mathcal{H}); (QH^2)^{\frac{1}{2}}\varepsilon) \leq C\varepsilon^{-2V/(V+2)},\ \varepsilon > 0,$$

where $C := C(K; V)$ (see (van der Vaart and Wellner, 1996)). One can easily compute in this case that

$$\int_0^x H^{1/2}_{d_{P_n,2}}(\mathcal{F}, u)du \leq \frac{1}{2}(V + 2)C^{1/2}(P_n H^2)^{\frac{V}{2(V+2)}} x^{\frac{2}{V+2}},\ x > 0 \text{ a.s.}$$

and, therefore, condition (2.2) of Theorem 1 is satisfied with $\psi(x) = x^{\frac{2}{V+2}}$ under the assumption $PH^2 < \infty$. Below we will assume that one of the two conditions holds:

1. Class $\mathcal{H}$ is uniformly bounded and $\mathcal{F} \subset \mathrm{conv}(\mathcal{H})$

2. The envelope H of the class $\mathcal{H}$ is $P-$square integrable and

$$\mathcal{F} \subset \Big\{\sum_{i=1}^N \lambda_i h_i : \ N \geq 1, h_i \in \mathcal{H}, \lambda_i \in \mathbb{R}, \sum_{j=1}^N |\lambda_j| = 1\Big\}.$$

Note, that under the second condition $\mathcal{F}$ consists only of proper symmetric convex combinations.

Let $\alpha := \frac{2V}{V+2}$ and $\Delta_f = \{\Delta \in [0,1] : d(f;\Delta) \leq n\}$. Define

$$\varepsilon_n(f;\delta) := \inf_{\Delta \in \Delta_f} \Big[\frac{d(f;\Delta)}{n}\Big(\log\frac{1}{\delta} + \log\frac{ne^2}{d(f;\Delta)}\Big) + \Big(\frac{\Delta}{\delta}\Big)^{\frac{2\alpha}{\alpha+2}} n^{-\frac{2}{\alpha+2}}\Big] \bigvee \frac{2\log n}{n}. \tag{2.10}$$

Let

$$\hat{\delta}_n(f) := \sup\Big\{\delta \in (0,1/2) : P_n\{f \leq \delta\} \leq \varepsilon_n(f;\delta)\Big\}.$$

Theorem 3 *Assume that one of the above conditions on the class $\mathcal{F}$ holds. Then there exist constants $A, B > 0$ such that for all $0 < t < n^{\frac{\alpha}{2+\alpha}}$ the following bound holds*

$$\mathbb{P}\Big\{\exists f \in \mathcal{F} \ P\{f \leq \frac{\hat{\delta}_n(f)}{4}\} \geq \ A\Big(\varepsilon_n(f;\frac{\hat{\delta}_n(f)}{2}) + \frac{t}{n}\Big)\Big\} \leq Be^{-t/4}.$$

Example 3. If $\mathcal{F} \subset \mathrm{conv}(\mathcal{H})$ is a class of functions such that for some $\beta > 0$

$$\sup_{f\in\mathcal{F}} d(f;\Delta) = O(\Delta^{-\beta}), \tag{2.11}$$

then with "high probability" for any classifier $f \in \mathcal{F}$ the upper bound on its generalization error becomes of the order

$$\frac{1}{n^{1-\gamma\beta/2(\gamma+\beta)}\hat{\delta}_n(f)^{\gamma\beta/(\gamma+\beta)}},$$

(which, of course, improves a more general bound in terms of γ-margins; the general bound corresponds to the case $\beta = +\infty$). The condition (2.11) means that the weights of the convex combination decrease polynomially fast, namely, $|\lambda_j| = O(j^{-\alpha})$, $\alpha = 1 + \beta^{-1}$. The case of exponential decrease of the weights is described by the condition

$$\sup_{f\in\mathcal{F}} d(f;\Delta) = O(\log\frac{1}{\Delta}). \tag{2.12}$$

In this case the upper bound becomes of the order $\frac{1}{n}\log^2\frac{n}{\hat{\delta}_n(f)}$.

3 Proofs of the main results

Proof of Theorem 1. We use the first bound of Theorem 2. The condition $\varepsilon \geq \varepsilon_n^\psi(\delta)$ is equivalent to the condition $\delta \geq \delta_n^\psi(\varepsilon)$. Thus, we can use this bound for $\delta = \delta_n^\psi(\varepsilon)$ and $\varepsilon \geq (2\log n)/n$. We get

$$\mathbb{P}\Big\{\exists f \in \mathcal{F}\ P_n\{f \leq \delta_n^\psi(\varepsilon)\} \leq \varepsilon \text{ and } P\{f \leq \frac{\delta_n^\psi(\varepsilon)}{2}\} \geq \bar{A}\varepsilon\Big\} \leq B\log_2\log_2\varepsilon^{-1}\exp\{-\frac{n\varepsilon}{2}\}.$$

Next we set $\varepsilon_j := 2^{-j}$. Let $\mathcal{J} = \{j \geq 0 : \varepsilon_j \geq \frac{t\vee 2\log n}{n}\}$ and

$$E := \Big\{\exists j \in \mathcal{J}\ \exists f \in \mathcal{F} :\ P_n\{f \leq \delta_n^\psi(\varepsilon_j)\} \leq \varepsilon_j \text{ and } P\{f \leq \frac{\delta_n^\psi(\varepsilon_j)}{2}\} \geq \bar{A}\varepsilon_j\Big\}.$$

We have

$$\mathbb{P}(E) \leq B\sum_{j\in\mathcal{J}}\log_2\log_2\varepsilon_j^{-1}\exp\Big\{-\frac{n\varepsilon_j}{2}\Big\} \leq B\log_2\log_2\frac{n}{t\vee 2\log n}\sum_{j\geq 0}\exp\Big\{-\big(\frac{t}{2}\vee\log n\big)2^j\Big\} \leq$$

$$\leq B'\log_2\log_2\frac{n}{t\vee 2\log n}\exp\Big\{-\big(\frac{t}{2}\vee\log n\big)\Big\}. \tag{3.1}$$

Suppose that for some j and for some $f \in \mathcal{F}$, $\hat{\varepsilon}_n^\psi(t;f) \in (\varepsilon_{j+1},\varepsilon_j]$. On the event E^c, the inequality $P_n\{f \leq \delta_n^\psi(\varepsilon_j)\} \leq \varepsilon_j$ implies that $P\{f \leq \delta_n^\psi(\varepsilon_j)/2\} \leq \bar{A}\varepsilon_j$. Since

$$\frac{\delta_n^\psi(\varepsilon_j)}{2} = \frac{\varphi^{-1}(\sqrt{\varepsilon_j n})}{2\sqrt{\varepsilon_j}} \geq \frac{\varphi^{-1}(\sqrt{4\varepsilon_j n})}{\sqrt{4\varepsilon_j}} = \delta_n^\psi(4\varepsilon_j),$$

we also have $P\{f \leq \delta_n^\psi(4\varepsilon_j)\} \leq \bar{A}\varepsilon_j$, which implies $P\{f \leq \delta_n^\psi(8\hat{\varepsilon}_n^\psi(f;t))\} \leq 2\bar{A}\hat{\varepsilon}_n^\psi(f;t)$. Therefore, on the event E^c, we get for all $f \in \mathcal{F}$, $\varepsilon_n^\psi(f;t) \leq (2\bar{A}\vee 8)\hat{\varepsilon}_n^\psi(f;t)$. It follows from (3.1) that

$$\mathbb{P}\Big\{\exists f \in \mathcal{F} :\ \varepsilon_n^\psi(f;t) \geq (2\bar{A}\vee 8)\hat{\varepsilon}_n^\psi(f;t)\Big\} \leq B'\log_2\log_2\frac{n}{t\vee 2\log n}\exp\Big\{-\big(\frac{t}{2}\vee\log n\big)\Big\}.$$

Quite similarly, using the second bound of Theorem 2, one can prove that

$$\mathbb{P}\Big\{\exists f \in \mathcal{F} :\ \hat{\varepsilon}_n^\psi(f;t) \geq (2\bar{A}\vee 8)\varepsilon_n^\psi(f;t)\Big\} \leq B'\log_2\log_2\frac{n}{t\vee 2\log n}\exp\Big\{-\big(\frac{t}{2}\vee\log n\big)\Big\},$$

which implies the inequality of Theorem 1.

□

Proof of Theorem 2. We define

$$r_0 := 1,\ r_{k+1} = C\sqrt{r_k \varepsilon} \bigwedge 1$$

where $C = c(1 + \mathbb{E}D_n)$ with a sufficiently large constant $c > 1$ (which will be chosen later). A simple induction shows that either $C\sqrt{\varepsilon} \geq 1$ and $r_k \equiv 1$, or $C\sqrt{\varepsilon} < 1$, and in the last case

$$r_k = C^{1+2^{-1}+\cdots+2^{-(k-1)}} \varepsilon^{2^{-1}+\cdots+2^{-k}} = C^{2(1-2^{-k})} \varepsilon^{1-2^{-k}} = (C\sqrt{\varepsilon})^{2(1-2^{-k})}.$$

Let $\gamma_k := (\varepsilon / r_k)^{1/2} = C^{2^{-k}-1} \varepsilon^{2^{-k-1}}$. Then

$$\begin{aligned} &\gamma_k + \gamma_{k-2} + \cdots + \gamma_0 = C^{-1}\big[C\sqrt{\varepsilon} + (C\sqrt{\varepsilon})^{2^{-1}} + \cdots + (C\sqrt{\varepsilon})^{2^{-k}}\big] \\ &\leq C^{-1}(C\sqrt{\varepsilon})^{2^{-k}}(1 - (C\sqrt{\varepsilon})^{2^{-k}})^{-1} \leq 1/2, \end{aligned}$$

for $\varepsilon \leq C^{-4}$, $C > 2(2^{1/4} - 1)^{-1}$ and $k \leq \log_2 \log_2 \varepsilon^{-1}$. For small enough ε (note that our choice of $\varepsilon \leq C^{-4}$ implies $C\sqrt{\varepsilon} < 1$), we have

$$\gamma_0 + \cdots + \gamma_k \leq \frac{1}{2},\ k \geq 1. \tag{3.2}$$

Let $\delta > 0$. Define

$$\delta_0 = \delta,\ \delta_k := \delta(1 - \gamma_0 - \ldots \gamma_{k-1}),\ \delta_{k,\frac{1}{2}} = \frac{1}{2}(\delta_k + \delta_{k+1}),\ \ k \geq 1.$$

Next we set $\mathcal{F}_0 := \mathcal{F}$, and define recursively

$$\mathcal{F}_{k+1} := \Big\{ f \in \mathcal{F}_k : P\{f \leq \delta_{k,\frac{1}{2}}\} \leq r_{k+1}/2 \Big\}.$$

For $k \geq 0$, let φ_k be a continuous function from $\mathbb{R}$ into $[0,1]$ such that $\varphi_k(u) = 1$ for $u \leq \delta_{k,\frac{1}{2}}$, $\varphi_k(u) = 0$ for $u \geq \delta_k$, and linear for $\delta_{k,\frac{1}{2}} \leq u \leq \delta_k$. For $k \geq 1$ let $\hat{\varphi}_k$ be a continuous function from $\mathbb{R}$ into $[0,1]$ such that $\hat{\varphi}_k(u) = 1$ for $u \leq \delta_k$, $\varphi_k(u) = 0$ for $u \geq \delta_{k-1,\frac{1}{2}}$, and linear for $\delta_k \leq u \leq \delta_{k-1,\frac{1}{2}}$. It follows from (3.2) that $\delta_k \in (\delta/2, \delta)$ for all $k \geq 1$. Note also that below our choice of k will be such that the restriction $k \leq \log_2 \log_2 \varepsilon^{-1}$ for any fixed $\varepsilon > 0$ will always be satisfied. Define

$$\mathcal{G}_k := \big\{ \varphi_k \circ f : f \in \mathcal{F}_k \big\},\ \ k \geq 0$$

and

$$\hat{\mathcal{G}}_k := \big\{ \hat{\varphi}_k \circ f : f \in \mathcal{F}_k \big\},\ \ k \geq 1.$$

It follows from the definitions that, for $k \geq 1$,

$$\sup_{g \in \hat{\mathcal{G}}_k} Pg^2 \leq \sup_{f \in \mathcal{F}_k} P\{f \leq \delta_k\} \leq \sup_{f \in \mathcal{F}_k} P\{f \leq \delta_{k-1,\frac{1}{2}}\} \leq r_k/2 \leq r_k$$

and

$$\sup_{g \in \mathcal{G}'_k} Pg^2 \leq \sup_{f \in \mathcal{F}_k} P\{f \leq \delta_{k-1,\frac{1}{2}}\} \leq r_k/2 \leq r_k.$$

Since $r_0 = 1$, for $k = 0$ the first inequality becomes trivial. Consider the following events

$$E^{(k)} := \Big\{ \|P_n - P\|_{\mathcal{G}_{k-1}} \le K_1 \mathbb{E}\|P_n - P\|_{\mathcal{G}_{k-1}} + K_2\sqrt{r_{k-1}\varepsilon} + K_3\varepsilon \Big\} \bigcap$$

$$\bigcap \Big\{ \|P_n - P\|_{\hat{\mathcal{G}}_k} \le K_1 \mathbb{E}\|P_n - P\|_{\hat{\mathcal{G}}_k} + K_2\sqrt{r_k\varepsilon} + K_3\varepsilon \Big\},\ k \ge 1,$$

By concentration inequalities of Talagrand (Talagrand, 1996b; Talagrand, 1996a) (see also (Massart, 1998)), for some values of numerical constants $K_1, K_2, K_3 > 0$,

$$\mathbb{P}((E^{(k)})^c) \le 2e^{-\frac{n\varepsilon}{2}}.$$

Denote $E_0 = \Omega$,

$$E_N := \bigcap_{k=1}^{N} E^{(k)},\ N \ge 1$$

and

$$\mathcal{J} = \Big\{ \inf_{f \in \mathcal{F}} P_n\{f \le \delta\} \le \varepsilon \Big\}.$$

Then

$$\mathbb{P}(E_N^c) \le 2Ne^{-\frac{n\varepsilon}{2}}.$$

In what follows we can and do assume without loss of generality that $\varepsilon < (2+C)^{-2}$ and therefore, $r_{k+1} < r_k$ and $\delta_k \in (\delta/2, \delta],\ k \ge 0$. [If $\varepsilon \ge (2+C)^{-2}$, then the bounds of the theorem obviously hold with any constant $A > 2 + C$.] The rest of the proof is based on the following lemma.

Lemma 1 *For any N such that*

$$N \le \log_2 \log_2 \varepsilon^{-1} \text{ and } r_N \ge \varepsilon, \tag{3.3}$$

the following properties hold on the event $E_N \bigcap \mathcal{J}$:

$$(i)\ \forall f \in \mathcal{F}\ P_n\{f \le \delta\} \le \varepsilon \Longrightarrow f \in \mathcal{F}_N$$

and

$$(ii)\ \sup_{f \in \mathcal{F}_k} P_n\{f \le \delta_k\} \le r_k,\ 0 \le k \le N.$$

Proof. We will prove the lemma by induction with respect to N. For $N = 0$, the statement is obvious. Suppose it holds for some $N \ge 0$, such that $N + 1$ still satisfies condition (3.3). Then, on the event $E_N \bigcap \mathcal{J}$,

$$\sup_{f \in \mathcal{F}_k} P_n\{f \le \delta_k\} \le r_k,\ 0 \le k \le N$$

and

$$\forall f \in \mathcal{F}\ P_n\{f \le \delta\} \le \varepsilon \Longrightarrow f \in \mathcal{F}_N.$$

Suppose that $f \in \mathcal{F}$ is such that $P_n\{f \le \delta\} \le \varepsilon$. By the induction assumptions, on the event E_N, we have $f \in \mathcal{F}_N$. Hence, on the event E_{N+1},

$$P\{f \le \delta_{N,\frac{1}{2}}\} \le P_n\{f \le \delta_N\} + \|P_n - P\|_{\mathcal{G}_N} \le$$
$$\le \varepsilon + K_1\mathbb{E}\|P_n - P\|_{\mathcal{G}_N} + K_2\sqrt{r_N\varepsilon} + K_3\varepsilon. \tag{3.4}$$

For a class $\mathcal{G}$, let

$$\hat{R}_n(\mathcal{G}) := \|n^{-1}\sum_{i=1}^n \varepsilon_i\delta_{X_i}\|_{\mathcal{G}},$$

where $\{\varepsilon_i\}$ is a sequence of i.i.d. Rademacher random variables.[1] By the symmetrization inequality,

$$\mathbb{E}\|P_n - P\|_{\mathcal{G}_N} \le 2\mathbb{E}I_{E_N}\mathbb{E}_\varepsilon\hat{R}_n(\mathcal{G}_N) + 2\mathbb{E}I_{E_N^c}\mathbb{E}_\varepsilon\hat{R}_n(\mathcal{G}_N). \tag{3.5}$$

Next, by the entropy inequalities for subgaussian processes (see (van der Vaart and Wellner, 1996), Corollary 2.2.8), we have

$$\mathbb{E}_\varepsilon\hat{R}_n(\mathcal{G}_N) \le \inf_{g\in\mathcal{G}_N}\mathbb{E}_\varepsilon\Big|n^{-1}\sum_{j=1}^n\varepsilon_j g(X_j)\Big| + \frac{\text{const}}{\sqrt{n}}\int_0^{(2\sup_{g\in\mathcal{G}_N}P_ng^2)^{1/2}} H^{1/2}_{d_{P_n,2}}(\mathcal{G}_N;u)du. \tag{3.6}$$

Remark. Here and in what follows in the proof "const" denotes a constant; its values can be different in different places.

By the induction assumption, on the event $E_N\bigcap\mathcal{J}$

$$\inf_{g\in\mathcal{G}_N}\mathbb{E}_\varepsilon\Big|n^{-1}\sum_{j=1}^n\varepsilon_j g(X_j)\Big| \le \inf_{g\in\mathcal{G}_N}\mathbb{E}_\varepsilon^{1/2}\Big|n^{-1}\sum_{j=1}^n\varepsilon_j g(X_j)\Big|^2 \le \frac{1}{\sqrt{n}}\inf_{g\in\mathcal{G}_N}\sqrt{P_ng^2} \le$$

$$\le \frac{1}{\sqrt{n}}\inf_{f\in\mathcal{F}_N}\sqrt{P_n\{f\le\delta_N\}} \le \frac{1}{\sqrt{n}}\inf_{f\in\mathcal{F}_N}\sqrt{P_n\{f\le\delta\}} \le \sqrt{\frac{\varepsilon}{n}} \le \varepsilon,$$

since $\varepsilon > n^{-1}$. We also have on the event $E_N\bigcap\mathcal{J}$

$$\sup_{g\in\mathcal{G}_N}P_ng^2 \le \sup_{f\in\mathcal{F}_N}P_n\{f\le\delta_N\} \le r_N.$$

The Lipschitz constants of φ_{k-1} and φ_k' is bounded by

$$L = 2(\delta_{k-1}-\delta_k)^{-1} = 2\delta^{-1}\gamma_{k-1}^{-1} = \frac{2}{\delta}\sqrt{\frac{r_{k-1}}{\varepsilon}}$$

which implies that

$$d^2_{P_n,2}\Big(\varphi_N\circ f;\varphi_N\circ g\Big) = n^{-1}\sum_{j=1}^n\Big|\varphi_N(f(X_j)) - \varphi_N(g(X_j))\Big|^2 \le \frac{2}{\delta}\sqrt{\frac{r_N}{\varepsilon}}d^2_{P_n,2}(f,g).$$

[1] The random variable $\hat{R}_n(\mathcal{G})$ is called *the Rademacher complexity* of the class $\mathcal{G}$. It was used by Koltchinskii (1999) (Koltchinskii, 1999), Bartlett, Boucheron and Lugosi (2000) (Bartlett et al., 2000), Koltchinskii and Panchenko (1999) (Koltchinskii and Panchenko, 2000) as a randomized complexity penalty in learning problems

It follows that, on the event $E_N \bigcap \mathcal{J}$,

$$\frac{1}{\sqrt{n}} \int_0^{(2\sup_{g\in\mathcal{G}_N} P_n g^2)^{1/2}} H^{1/2}_{d_{P_n,2}}(\mathcal{G}_N; u)du \le \frac{1}{\sqrt{n}} \int_0^{(2r_N)^{1/2}} H^{1/2}_{d_{P_n,2}}(\mathcal{F}; \frac{\delta\sqrt{\varepsilon}u}{2\sqrt{r_N}})du$$
$$\le \frac{1}{\sqrt{n}} \frac{2\sqrt{r_N}}{\delta\sqrt{\varepsilon}} \int_0^{\delta\sqrt{\varepsilon}/2} H^{1/2}_{d_{P_n,2}}(\mathcal{F}; v)dv \le \frac{1}{\sqrt{n}} \frac{2\sqrt{r_N}}{\delta\sqrt{\varepsilon}} D_n \psi\big(\frac{\delta\sqrt{\varepsilon}}{2}\big) \le$$
$$\frac{2D_n\sqrt{r_N}}{\sqrt{\varepsilon}}\varepsilon = 2D_n\sqrt{r_N\varepsilon}, \tag{3.7}$$

where we used the fact that for $\varepsilon \ge \varepsilon_n^{\psi}(\delta)$ the inequality $\psi(\delta\sqrt{\varepsilon}/2)/(\delta\sqrt{n}) \le \varepsilon$ holds. Now (3.6) and (3.7) imply that on the event $E_{N+1} \bigcap \mathcal{J}$

$$\mathbb{E}_\varepsilon \hat{R}_n(\mathcal{G}_N) \le \mathrm{const}(1 + D_n)\sqrt{r_N\varepsilon}. \tag{3.8}$$

Since we also have $\mathbb{E}_\varepsilon \hat{R}_n(\mathcal{G}_{N+1}) \le 1$, (3.5) and (3.8) yield

$$\mathbb{E}\|P_n - P\|_{\mathcal{G}_N} \le \mathrm{const}(1 + \mathbb{E}D_n)\sqrt{r_N\varepsilon} + 2\mathbb{P}(E_N^c) \le \mathrm{const}(1 + \mathbb{E}D_n)\sqrt{r_N\varepsilon} + 4Ne^{-n\varepsilon/2}.$$

By condition (3.3) and the fact that $\varepsilon \ge 2\log n/n$, we have $4Ne^{-n\varepsilon/2} \le \varepsilon$. Therefore,

$$\mathbb{E}\|P_n - P\|_{\mathcal{G}_N} \le \mathrm{const}(1 + \mathbb{E}D_n)\sqrt{r_N\varepsilon}.$$

By (3.4), on the event $E_{N+1} \bigcap \mathcal{J}$

$$P\{f \le \delta_{N,\frac{1}{2}}\} \le \mathrm{const}(1 + \mathbb{E}D_n)\big(\varepsilon + \sqrt{r_N\varepsilon}\big). \tag{3.9}$$

It follows that with a proper choice of constant $c > 0$ in the recurrent relationship defining the sequence $\{r_k\}$, we have on the event $E_{N+1} \bigcap \mathcal{J}$

$$P\{f \le \delta_{N,\frac{1}{2}}\} \le \frac{1}{2}C\sqrt{r_N\varepsilon} = r_{N+1}/2.$$

This means that $f \in \mathcal{F}_{N+1}$ and the induction step for (i) is proved.

To prove (ii), note that on the event E_{N+1}

$$\sup_{f\in\mathcal{F}_{N+1}} P_n\{f \le \delta_{N+1}\} \le \sup_{f\in\mathcal{F}_{N+1}} P\{f \le \delta_{N,\frac{1}{2}}\} + \|P_n - P\|_{\hat{\mathcal{G}}_{N+1}} \le$$
$$\le r_{N+1}/2 + K_1\mathbb{E}\|P_n - P\|_{\hat{\mathcal{G}}_{N+1}} + K_2\sqrt{r_{N+1}\varepsilon} + K_3\varepsilon. \tag{3.10}$$

By the symmetrization inequality,

$$\mathbb{E}\|P_n - P\|_{\hat{\mathcal{G}}_{N+1}} \le 2\mathbb{E}I_{E_N}\mathbb{E}_\varepsilon \hat{R}_n(\hat{\mathcal{G}}_{N+1}) + 2\mathbb{E}I_{E_N^c}\mathbb{E}_\varepsilon \hat{R}_n(\hat{\mathcal{G}}_{N+1}). \tag{3.11}$$

Similarly to (3.6)

$$\mathbb{E}_\varepsilon R_n(\hat{\mathcal{G}}_{N+1}) \le \inf_{g\in\hat{\mathcal{G}}_{N+1}} \mathbb{E}_\varepsilon\big|n^{-1}\sum_{j=1}^n \varepsilon_j g(X_j)\big| + \frac{\mathrm{const}}{\sqrt{n}} \int_0^{(2\sup_{g\in\hat{\mathcal{G}}_{N+1}} P_n g^2)^{1/2}} H^{1/2}_{d_{P_n,2}}(\hat{\mathcal{G}}_{N+1}; u)du. \tag{3.12}$$

It follows from (i) that on the event $E_{N+1}\bigcap\mathcal{J}$

$$\inf_{g\in\hat{\mathcal{G}}_{N+1}}\mathbb{E}_\varepsilon\big|n^{-1}\sum_{j=1}^n\varepsilon_j g(X_j)\big|\le\inf_{g\in\hat{\mathcal{G}}_{N+1}}\mathbb{E}_\varepsilon^{1/2}\big|n^{-1}\sum_{j=1}^n\varepsilon_j g(X_j)\big|^2\le\frac{1}{\sqrt{n}}\inf_{g\in\hat{\mathcal{G}}_{N+1}}\sqrt{P_ng^2}\le$$

$$\le\frac{1}{\sqrt{n}}\inf_{f\in\mathcal{F}_{N+1}}\sqrt{P_n\{f\le\delta_{N,\frac12}\}}\le\frac{1}{\sqrt{n}}\inf_{f\in\mathcal{F}_{N+1}}\sqrt{P_n\{f\le\delta\}}\le\sqrt{\frac{\varepsilon}{n}}\le\varepsilon.$$

By the induction assumption, we also have on the event $E_{N+1}\bigcap\mathcal{J}$

$$\sup_{g\in\hat{\mathcal{G}}_{N+1}}P_ng^2\le\sup_{f\in\mathcal{F}_N}P_n\{f\le\delta_{N,\frac12}\}\le r_N.$$

The bound for the Lipschitz constant of φ_k' yields

$$d^2_{P_n,2}\Big(\hat\varphi_{N+1}\circ f;\hat\varphi_{N+1}\circ g\Big)=n^{-1}\sum_{j=1}^n\Big|\hat\varphi_{N+1}\circ f(X_j)-\hat\varphi_{N+1}\circ g(X_j)\Big|^2\le\frac{2}{\delta}\sqrt{\frac{r_N}{\varepsilon}}d^2_{P_n,2}(f,g).$$

Hence, on the event $E_{N+1}\bigcap\mathcal{J}$, we get quite similarly to (3.7)

$$\frac{1}{\sqrt{n}}\int_0^{(2\sup_{g\in\hat{\mathcal{G}}_{N+1}}P_ng^2)^{1/2}}H^{1/2}_{d_{P_n,2}}(\mathcal{G}'_{N+1};u)du\le\frac{1}{\sqrt{n}}\int_0^{(2r_N)^{1/2}}H^{1/2}_{d_{P_n,2}}(\mathcal{F};\frac{\delta\sqrt{\varepsilon}u}{2\sqrt{r_N}})du$$

$$\le\frac{1}{\sqrt{n}}\frac{2\sqrt{r_N}}{\delta\sqrt{\varepsilon}}\int_0^{\delta\sqrt{\varepsilon}/2}H^{1/2}_{d_{P_n,2}}(\mathcal{F};v)dv\le\frac{1}{\sqrt{n}}\frac{2\sqrt{r_N}}{\delta\sqrt{\varepsilon}}D_n\psi(\frac{\delta\sqrt{\varepsilon}}{2})\le$$

$$\frac{2D_n\sqrt{r_N}}{\sqrt{\varepsilon}}\varepsilon=2D_n\sqrt{r_N\varepsilon}.\tag{3.13}$$

Combining all the bounds, we get on the event $E_{N+1}\bigcap\mathcal{J}$

$$\sup_{f\in\mathcal{F}_{N+1}}P_n\{f\le\delta_{N+1}\}\le\frac{r_{N+1}}{2}+\mathrm{const}(1+\mathbb{E}D_n)\sqrt{r_N\varepsilon}.\tag{3.14}$$

With a proper choice of constant $c>0$ in the recurrent relationship defining the sequence $\{r_k\}$, we have on the event $E_{N+1}\bigcap\mathcal{J}$

$$\sup_{f\in\mathcal{F}_{N+1}}P_n\{f\le\delta_{N+1}\}\le C\sqrt{r_N\varepsilon}=r_{N+1},$$

which proves the induction step for (ii) and the lemma.

□

To complete the proof of the theorem, note that the choice of $N=[\log_2\log_2\varepsilon^{-1}]$ implies that $r_{N+1}\le c\varepsilon$ for some $c>0$. Indeed, if we introduce $s_k=r_k/C$ and $\varepsilon_1=C\varepsilon$ then $s_{k+1}=\sqrt{s_k\varepsilon}$ and $s_0=C^{-1}\le1$. It is easy to see that $s_N\le\varepsilon_1^{1-2^{-N}}\le2\varepsilon_1$ for $N\ge\log_2\log_2\varepsilon_1^{-1}$, and, hence, $r_N\le C^2\varepsilon=\bar A\varepsilon$.

The proof of the second inequality is similar with minor modifications.

□

To prove Theorem 3, we need the following statement, which seems to be well known, but we have not found the precise reference and give the proof here for completeness.

Let

$$\mathrm{conv}_d(\mathcal{H}) := \Big\{\sum_{j=1}^{d} \lambda_j h_j : \lambda_j \in \mathbb{R}, \sum_{j=1}^{d} |\lambda_j| \le 1,\ h_j \in \mathcal{H}\Big\}.$$

Lemma 2 *Let $\mathcal{H}$ be a class of functions from $(S, \mathcal{A})$ into $\mathbb{R}$. Let Q be a probability measure on $(S, \mathcal{A})$ such that*

$$\bar{H} := \sup_{h\in\mathcal{H}} (Qh^2)^{1/2} < +\infty.$$

The following bound holds for all $d \ge 1$ and $\varepsilon > 0$:

$$N_{d_{Q,2}}\Big(\mathrm{conv}_d(\mathcal{H}), (1+\bar{H})\varepsilon\Big) \le \left(\frac{2e^2 N_{d_{Q,2}}(\mathcal{H},\varepsilon)(d' + 4\varepsilon^{-2})}{d'^2}\right)^{d'},$$

where $d' = d \wedge N_{d_{Q,2}}(\mathcal{H}, \varepsilon)$.

Proof. First note that if $\mathcal{H}' := \mathcal{H} \bigcup \big\{h : -h \in \mathcal{H}\big\}$, then $\mathrm{conv}_d(\mathcal{H}') = \mathrm{conv}_d(\mathcal{H})$ and

$$N_{d_{Q,2}}(\mathcal{H}';\varepsilon) \le 2N_{d_{Q,2}}(\mathcal{H};\varepsilon).$$

Thus, it's enough to show that for a class $\mathcal{H}$, such that $h \in \mathcal{H}$ implies $-h \in \mathcal{H}$, we have

$$N_{d_{Q,2}}\Big(\mathrm{conv}_d(\mathcal{H}), (1+\bar{H})\varepsilon\Big) \le \left(\frac{e^2 N_{d_{Q,2}}(\mathcal{H},\varepsilon)(d + 4\varepsilon^{-2})}{d^2}\right)^{d}.$$

For such a class we have

$$\mathrm{conv}_d(\mathcal{H}) := \Big\{\sum_{j=1}^{d} \lambda_j h_j : \lambda_j \ge 0, \sum_{j=1}^{d} \lambda_j \le 1,\ h_j \in \mathcal{H}\Big\}.$$

Note that if $\sum_j |\lambda_j| \le 1$, then

$$d_{Q,2}\Big(\sum_j \lambda_j h_j; \sum_j \lambda_j h'_j\Big) = \Big\|\sum_j \lambda_j (h_j - h'_j)\Big\|_{L_2(Q)} \le$$

$$\le \sum_j |\lambda_j| \max_j \big\|h_j - h'_j\big\|_{L_2(Q)} \le \max_j \big\|h_j - h'_j\big\|_{L_2(Q)}.$$

It follows that if $\mathcal{H}_\varepsilon$ is an ε-net of $\mathcal{H}$, then a δ-net of $\mathrm{conv}_d(\mathcal{H}_\varepsilon)$ is an $\varepsilon + \delta$-net of $\mathrm{conv}_d(\mathcal{H})$. This observation allows us to reduce the proof of the lemma to the case when $\mathcal{H}$ is a finite class. In this case we want to show that

$$N_{d_{Q,2}}\Big(\mathrm{conv}_d(\mathcal{H}), \bar{H}\varepsilon\Big) \le \left(\frac{e^2 \mathrm{card}(\mathcal{H})(d + 4\varepsilon^{-2})}{d^2}\right)^{d}.$$

To this end, we use the idea of B. Maurey (Pisier, 1981; van der Vaart and Wellner, 1996). Let $N := \mathrm{card}(\mathcal{H})$. Consider some representation of a function $f = \sum_{i=1}^N \lambda_i h_i \in \mathrm{conv}_d(\mathcal{H})$. We assume that $\lambda_j \geq 0$, $\sum_j \lambda_j \leq 1$, and at most d' of the coefficients are not equal to 0. Consider an i.i.d. sequence of random variables Y_j, $j = 1, \dots, k$ taking values in $\mathcal{H} \cup \{0\}$ such that $P(Y_j = h_i) = \lambda_i$ for $i = 1, \dots, N$ and $P(Y_j = 0) = 1 - \sum_{i=1}^N \lambda_i$. (We simply add the probabilities when the same function h corresponds to several weights λ_i with different indices). We have

$$\mathbb{E}\|k^{-1}\sum_{j=1}^k Y_j - \sum_{i=1}^N \lambda_i h_i\|_{Q,2}^2 = \mathbb{E}\|k^{-1}\sum_{j=1}^k Y_j - \mathbb{E}Y_1\|_{Q,2}^2 \leq$$
$$\leq \frac{1}{k}\mathbb{E}\|Y_1 - \mathbb{E}Y_1\|_{Q,2}^2 \leq 4\bar{H}^2 k^{-1}.$$

If we set $k = 4\varepsilon^{-2}$, then with probability 1 there exists a realization $\bar{Y}_k = k^{-1}\sum_{j=1}^k Y_j$ such that

$$\|\bar{Y}_k - \sum_{i=1}^N \lambda_i h_i\|_{Q,2} \leq \varepsilon \bar{H}.$$

In order to compute the bound for the $\bar{H}\varepsilon-$covering number we have to calculate the number of possible realizations of $k^{-1}\sum_{j=1}^k Y_j$. A simple combinatorics shows that this number does not exceed $\binom{N}{d'}\binom{d'+k}{k}$. Next we use the following bound, which holds for all $1 \leq d \leq N$:

$$\binom{N}{d}\binom{d+k}{k} \leq \left(\frac{e^2 N(d+k)}{d^2}\right)^d.$$

To prove the bound, first assume that $d < N$. Then one can check using Stirling's formula that

$$\frac{N!}{d!(N-d)!}\frac{(d+k)!}{d!k!} \leq \frac{n^n}{d^d(N-d)^{N-d}}\frac{(d+k)^{d+k}}{k^k d^d}$$
$$\leq \left(\frac{N(d+k)}{d^2}\right)^d \left(1 + \frac{d}{N-d}\right)^{N-d}\left(1 + \frac{d}{k}\right)^k \leq \left(\frac{e^2 N(d+k)}{d^2}\right)^d.$$

The case when $d = N$ can be considered similarly. The bound immediately implies the result.

□

Proof of Theorem 3. Let us fix $\delta \in (0, 1/2]$. For any function f we denote $d(f) := d(f, \bar{\Delta})$, where $\bar{\Delta}$ is such that the infimum in the definition (2.10) is attained at $\bar{\Delta}$. For a fixed δ we consider a partition of $\mathcal{F}$ into two classes $\mathcal{F}_1^\delta$ and $\mathcal{F}_2^\delta = \mathcal{F} \setminus \mathcal{F}_1^\delta$, where $\mathcal{F}_1^\delta := \{f : d(f) = 0\}$ (note that $d(f)$ depends on δ). In the first four steps of the proof we will deal with $\mathcal{F}_2^\delta$ and we will assume only that the class $\mathcal{H}$ has a square integrable envelope H.

Step 1. Let $1 \leq d \leq n$. Denote

$$\varepsilon_n(d; \delta; \Delta) := \Big[\frac{d}{n}\Big(\log\frac{1}{\delta} + \log\frac{ne^2}{d}\Big) + \Big(\frac{\Delta}{\delta}\Big)^{\frac{2\alpha}{\alpha+2}} n^{-\frac{2}{\alpha+2}}\Big] \bigvee \frac{2\log n}{n}.$$

Let $\mathcal{F}_{d,\Delta} := \{f \in \mathcal{F}_2^\delta : d(f;\Delta) \le d\}$. We start by proving (with some constants $A, B > 0$) the following inequality:

$$\mathbb{P}\Big\{\exists f \in \mathcal{F}_{d,\Delta}\ P_n\{f \le \delta\} \le \varepsilon_n(d;\delta;\Delta) \text{ and } P\{f \le \frac{\delta}{2}\} \ge A\varepsilon_n(d;\delta;\Delta)\Big\} \le$$
$$\le B\Big(\frac{\delta d}{n}\Big)^{d/4} \exp\Big\{-\frac{1}{4}\big(\sqrt{n}\frac{\Delta}{\delta}\big)^{2\alpha/(\alpha+2)}\Big\}. \tag{3.15}$$

Clearly, we can and do assume that $\varepsilon_n(d;\delta;\Delta) \le 1$. To prove (3.15), we bound the random entropy $H_{d_{P_n,2}}(\mathcal{F}_{d,\Delta};\varepsilon)$ of the class $\mathcal{F}_{d,\Delta}$ the following way:

$$H_{d_{P_n,2}}(\mathcal{F}_{d,\Delta};\varepsilon) \le K(1+P_nH^2)\Big[d\log\frac{e}{\varepsilon} + \Big(\frac{\Delta}{\varepsilon}\Big)^\alpha\Big] \text{ for } \varepsilon \le 1 \tag{3.16}$$

with some constant $K > 0$. The last bound follows from the observation that each function $f \in \mathcal{F}_{d,\Delta}$ can be represented as $f = f_1 + f_2$, where

$$f_1 \in \mathcal{F}_d := \mathrm{conv}_d(\mathcal{H}) = \Big\{\sum_{j=1}^d \lambda_j h_j : \lambda_j \in \mathbb{R}, \sum_{j=1}^d |\lambda_j| \le 1,\ h_j \in \mathcal{H}\Big\}$$

and

$$f_2 \in \mathcal{F}_\Delta := \Delta\ \mathrm{conv}(\mathcal{H}).$$

Hence, by simple combining of ε-coverings for the classes $\mathcal{F}_d$ and $\mathcal{F}_\Delta$, we get

$$H_{d_{P_n,2}}(\mathcal{F}_{d,\Delta};\varepsilon) \le H_{d_{P_n,2}}(\mathcal{F}_d;\varepsilon/2) + H_{d_{P_n,2}}(\mathcal{F}_\Delta;\varepsilon/2).$$

Then, a routine application of Lemma 2 and (2.9) implies

$$H_{d_{P_n,2}}(\mathcal{F}_d;\varepsilon/2) \le Kd\log\frac{e(1+P_nH^2)}{\varepsilon} \text{ for } \varepsilon \le 2(P_nH^2)^{1/2}$$

(note that for $\varepsilon > 2(P_nH^2)^{1/2}$ we easily get $H_{d_{P_n,2}}(\mathcal{F}_d;\varepsilon/2) = 0$). For $\varepsilon \le 1$ this implies

$$H_{d_{P_n,2}}(\mathcal{F}_d;\varepsilon/2) \le Kd\Big[\log\frac{e}{\varepsilon} + \log(1+P_nH^2)\Big] \le Kd\Big[\log\frac{e}{\varepsilon} + P_nH^2\Big] \le Kd(1+P_nH^2)\log\frac{e}{\varepsilon}.$$

By the bound on the entropy of the symmetric convex hull (van der Vaart and Wellner, 1996)

$$H_{d_{P_n,2}}(\mathcal{F}_\Delta;\varepsilon/2) = H_{d_{P_n,2}}\big(\mathcal{F};\frac{\varepsilon}{2\Delta}\big) \le K(1+P_nH^2)^{\alpha/4}\Big(\frac{\Delta}{\varepsilon}\Big)^\alpha \le K(1+P_nH^2)\Big(\frac{\Delta}{\varepsilon}\Big)^\alpha,$$

which implies (3.16).

Next we are using margin-type bounds on generalization error under random entropy conditions (see Section 2, Theorem 2). Clearly, from (3.16), we get the following bound on Dudley's entropy integral:

$$\int_0^x H^{1/2}_{d_{P_n,2}}(\mathcal{F};\varepsilon)d\varepsilon \le K(1+P_nH^2)^{1/2}\bar{\psi}(x),$$

where $\bar{\psi}$ is a concave nondecreasing function such that for $x \in [0,1]$

$$\bar{\psi}(x) = \Big(x \Big(d \log \frac{e}{x} \Big)^{1/2} + \Delta^{\alpha/2} x^{1-\alpha/2} \Big)$$

with some constant $K > 0$. Let

$$\psi_1(x) := x \Big(d \log \frac{e}{x} \Big)^{1/2}, \ \psi_2(x) := \Delta^{\alpha/2} x^{1-\alpha/2}, \ \psi(x) := (\psi_1(x) + \psi_2(x))/2.$$

Let us first consider the equation $\varepsilon = \psi_1(\delta\sqrt{\varepsilon})/(\delta\sqrt{n})$, which can be written as $\varepsilon = \frac{d}{n} \log \frac{e}{\delta\sqrt{\varepsilon}}$. If $\varepsilon = \frac{d}{n} x^2$ then

$$x e^{x^2} = \Big(\frac{n}{d} \Big)^{1/2} \frac{e}{\delta}.$$

For $d \le n$ and $\delta \le 1$, it means that $x e^{x^2} \ge 1$, and, therefore,

$$e^{x^2 - 1} \le \Big(\frac{n}{d} \Big)^{1/2} \frac{e}{\delta},$$

or,

$$\varepsilon = \frac{d}{n} x^2 \le \frac{d}{n} \Big[1 + \log \Big(\Big(\frac{n}{d} \Big)^{1/2} \frac{e}{\delta} \Big) \Big] \le \frac{d}{n} \log \frac{n e^2}{d\delta} \le \varepsilon_n(d; \delta; \Delta) \le 1.$$

[One can notice that in the case when d becomes significantly greater then n, for example, if $(nd^{-1})^{1/2} \delta^{-1} \le 1$ then $x \le 1$ and $x e^{x^2} \le ex$, which implies that $\varepsilon \ge \delta^{-2}$ and the bound of the theorem becomes useless. This explains why in the definition of $\varepsilon_n(f; \delta)$ we minimize over $d(f, \Delta) \le n$.]

The solution of the equation $\varepsilon = \psi_2(\delta\sqrt{\varepsilon})/(\delta\sqrt{n})$ is equal to

$$\varepsilon^{(2)} := \Big(\frac{\Delta}{\delta} \Big)^{\frac{2\alpha}{\alpha+2}} n^{-\frac{2}{\alpha+2}}.$$

Finally, it is easy to bound the solution of the equation $\varepsilon = \psi(\delta\sqrt{\varepsilon})/(\delta\sqrt{n})$ from above by $\varepsilon^{(1)} + \varepsilon^{(2)}$. Therefore, the solution of the last equation is also bounded from above by $\varepsilon_n(d; \delta; \Delta)$. This allows us to use the bound of Theorem 2 to get the following inequality:

$$\mathbb{P} \Big\{ \exists f \in \mathcal{F}_{d,\Delta} \ P_n\{f \le \delta\} \le \varepsilon_n(d; \delta; \Delta) \text{ and } P\{f \le \frac{\delta}{2}\} \ge A \varepsilon_n(d; \delta; \Delta) \Big\} \le$$

$$\le B \log_2 \log_2 \varepsilon_n(d; \delta; \Delta)^{-1} \exp\{ - \frac{n \varepsilon_n(d; \delta; \Delta)}{2} \}.$$

Since, for $\varepsilon := \varepsilon_n(d; \delta; \Delta)$, we have $\varepsilon \ge \frac{2 \log n}{n}$, it follows that for $n \ge 3$ (??)

$$\frac{1}{\varepsilon} \log \log_2 \log_2 \frac{1}{\varepsilon} \le n/4,$$

which implies

$$B \log_2 \log_2 \varepsilon_n(d; \delta; \Delta)^{-1} \exp\{ - \frac{n \varepsilon_n(d; \delta; \Delta)}{2} \} \le B \exp\{ - \frac{n \varepsilon_n(d; \delta; \Delta)}{4} \}. \tag{3.17}$$

A simple computation shows that

$$\exp\{-\frac{n\varepsilon_n(d;\delta;\Delta)}{4}\} \le \Big(\frac{\delta d}{n}\Big)^{d/4} \exp\Big\{-\frac{1}{4}\big(\sqrt{n}\frac{\Delta}{\delta}\big)^{2\alpha/(\alpha+2)}\Big\},$$

which implies (3.15)

Step 2. Next we show that with some constants $A, B \ge 1$, $\delta \le 1/2$ and $\Delta \ge \delta n^{-1/2}$

$$\mathbb{P}\Big\{\exists f \in \mathcal{F}_2^\delta \; P_n\{f \le \delta\} \le \varepsilon_n(d(f;\Delta);\delta;\Delta) \text{ and } P\{f \le \frac{\delta}{2}\} \ge A\varepsilon_n(d(f;\Delta);\delta;\Delta)\Big\} \le$$
$$\le B\delta^{1/8}\Delta^{1/8} \exp\Big\{-\frac{1}{4}\big(\sqrt{n}\frac{\Delta}{\delta}\big)^{2\alpha/(\alpha+2)}\Big\}, \tag{3.18}$$

where it's understood that if $d = d(f;\Delta) > n$ then $\varepsilon_n(d;\delta;\Delta) = 1$. Indeed, using (3.15), we have for $\delta \le 1/2$

$$\mathbb{P}\Big\{\exists f \in \mathcal{F}_2^\delta \; P_n\{f \le \delta\} \le \varepsilon_n(d(f;\Delta);\delta;\Delta) \text{ and } P\{f \le \frac{\delta}{2}\} \ge A\varepsilon_n(d(f;\Delta);\delta;\Delta)\Big\} \le$$

$$\le \mathbb{P}\Big\{\exists d \le n \; \exists f \in \mathcal{F}_2^\delta \; d(f;\Delta) = d, P_n\{f \le \delta\} \le \varepsilon_n(d;\delta;\Delta) \text{ and } P\{f \le \frac{\delta}{2}\} \ge A\varepsilon_n(d;\delta;\Delta)\Big\} \le$$

$$\le \sum_{d=1}^n \mathbb{P}\Big\{\exists f \in \mathcal{F}_{d,\Delta} \; P_n\{f \le \delta\} \le \varepsilon_n(d;\delta;\Delta) \text{ and } P\{f \le \frac{\delta}{2}\} \ge A\varepsilon_n(d;\delta;\Delta)\Big\} \le$$

$$\le B\sum_{d=1}^n \Big(\frac{\delta d}{n}\Big)^{d/4} \exp\Big\{-\frac{1}{4}\big(\sqrt{n}\frac{\Delta}{\delta}\big)^{2\alpha/(\alpha+2)}\Big\}.$$

One can easily check that for $d \le n/(e\delta)$ (increasing A we can assume that it holds) the expression $(\delta d/n)^{d/4}$ is decreasing in d and, therefore, for any $k \le n/e$

$$\sum_{d=1}^n \Big(\frac{\delta d}{n}\Big)^{d/4} \le k\Big(\frac{\delta}{n}\Big)^{1/4} + \sum_{d=k+1}^n \Big(\frac{\delta d}{n}\Big)^{d/4} \le k\Big(\frac{\delta}{n}\Big)^{1/4} + \delta^{k/4}.$$

Optimizing over k we take $k = \log n/\log\delta^{-1} + 1$ to get

$$k\Big(\frac{\delta}{n}\Big)^{1/4} + \delta^{k/4} \le 2\Big(\frac{\log n}{\log\delta^{-1}} + 1\Big)\Big(\frac{\delta}{n}\Big)^{1/4} \le \delta^{1/8}\Delta^{1/8},$$

where the last inequality holds under the assumption that $\Delta \ge \delta n^{-1/2}$.

Step 3. Our next goal is to prove that with some constants $A, B > 1$ and for $0 < t < n^{\alpha/(2+\alpha)}$

$$\mathbb{P}\Big\{\exists f \in \mathcal{F}_2^\delta \; P_n\{f \le \delta\} \le \varepsilon_n(f;\delta) \text{ and } P\{f \le \frac{\delta}{2}\} \ge A \inf_{\Delta \ge \delta n^{-1/2} t^{\frac{1}{\alpha}+\frac{1}{2}}} \varepsilon_n(d(f;\Delta);\delta;\Delta)\Big\} \le$$
$$\le B\delta^{1/8}e^{-t/4} \tag{3.19}$$

Let $\Delta_j := 2^{-j},\ j \geq 0$. Let $\mathcal{J} = \{j : \Delta_j \geq \delta n^{-1/2} t^{\frac{1}{\alpha}+\frac{1}{2}}\}$. Note that the condition $t < n^{\alpha/(2+\alpha)}$ guarantees that $\mathcal{J} \neq \emptyset$. Using (3.18), we get

$$\mathbb{P}\Big\{\exists f \in \mathcal{F}_2^\delta\ P_n\{f \leq \delta\} \leq \varepsilon_n(f;\delta) \text{ and } P\{f \leq \frac{\delta}{2}\} \geq A \inf_{\mathcal{J}} \varepsilon_n(d(f;\Delta_j);\delta;\Delta_j)\Big\} \leq$$
$$\leq \mathbb{P}\Big\{\exists f \in \mathcal{F}_2^\delta\ \exists j \in \mathcal{J}\ P_n\{f \leq \delta\} \leq \varepsilon_n(f;\delta) \text{ and } P\{f \leq \frac{\delta}{2}\} \geq A\varepsilon_n(d(f;\Delta_j);\delta;\Delta_j)\Big\} \leq$$
$$\leq \sum_{\mathcal{J}} \mathbb{P}\Big\{\exists f \in \mathcal{F}_2^\delta\ P_n\{f \leq \delta\} \leq \varepsilon_n(f;\delta) \text{ and } P\{f \leq \frac{\delta}{2}\} \geq A\varepsilon_n(d(f;\Delta_j);\delta;\Delta_j)\Big\} \leq$$
$$\leq B \sum_{\mathcal{J}} \delta^{1/8} \Delta_j^{1/8} \exp\Big\{-\frac{1}{4}\big(\sqrt{n}\frac{\Delta_j}{\delta}\big)^{2\alpha/(\alpha+2)}\Big\} \leq B'\delta^{1/8} e^{-t/4}.$$

To complete the proof of (3.19), note that for $\Delta \in (\Delta_{j+1}, \Delta_j]$ we have

$$\frac{d(f;\Delta_j)}{n}\Big(\log\frac{1}{\delta} + \log\frac{ne}{d(f;\Delta_j)}\Big) \leq \frac{d(f;\Delta)}{n}\Big(\log\frac{1}{\delta} + \log\frac{ne}{d(f;\Delta)}\Big),$$

$$\Big(\frac{\Delta_j}{\delta}\Big)^{\frac{2\alpha}{\alpha+2}} n^{-\frac{2}{\alpha+2}} \leq 2^{\frac{2\alpha}{(\alpha+2)}}\Big(\frac{\Delta}{\delta}\Big)^{\frac{2\alpha}{\alpha+2}} n^{-\frac{2}{\alpha+2}},\ \ \log\log\frac{2}{\Delta_j} \leq \log\log\frac{2}{\Delta},$$

which implies $\varepsilon_n(f;\Delta_j;\delta) \leq 2^{2\alpha/(\alpha+2)}\varepsilon_n(f;\Delta;\delta)$ and, therefore,

$$\inf_{\mathcal{J}} \varepsilon_n(d(f;\Delta_j);\delta;\Delta_j) \leq 2^{2\alpha/(\alpha+2)} \inf_{\Delta \geq \delta n^{-1/2} t^{\frac{1}{\alpha}+\frac{1}{2}}} \varepsilon_n(d(f;\Delta);\delta;\Delta),$$

and (3.19) follows.

Step 4. Now we prove that for some constants $A, B > 1$ and for all $0 < t < n^{\alpha/2+\alpha}$

$$\mathbb{P}\Big\{\exists f \in \mathcal{F}_2^\delta\ P_n\{f \leq \delta\} \leq \varepsilon_n(f;\delta) \text{ and } P\{f \leq \frac{\delta}{2}\} \geq A\big(\varepsilon_n(f;\delta) + \frac{t}{n}\big)\Big\} \leq$$
$$\leq B\delta^{1/8} e^{-t/4} \tag{3.20}$$

Because of (3.19), it is enough to show that

$$\inf_{\Delta \geq \delta n^{-1/2} t^{\frac{1}{\alpha}+\frac{1}{2}}, \Delta \in \Delta_f} \varepsilon_n(d(f;\Delta);\delta;\Delta) \leq \varepsilon_n(f;\delta) + \frac{t}{n}. \tag{3.21}$$

Since $d(f;\Delta)$ is a decreasing function of Δ, the set Δ_f is an interval of the form $[c,1]$ for some $c \leq 1$. Let $\Delta_0 := \delta n^{-1/2} t^{\frac{1}{\alpha}+\frac{1}{2}}$. If $\Delta_0 \notin \Delta_f$, then (3.21) clearly holds. Otherwise, suppose that the infimum in the definition of $\varepsilon_n(f;\delta)$ is attained at $\Delta = \bar{\Delta}$. If $\bar{\Delta} \geq \Delta_0$, then (3.21) is also obvious. In the case when $\bar{\Delta} < \Delta_0$, note that

$$\Big(\frac{\Delta_0}{\delta}\Big)^{\frac{2\alpha}{\alpha+2}} n^{-\frac{2}{\alpha+2}} = \frac{t}{n}$$

and the function $\frac{d(f;\Delta)}{n}\Big(\log\frac{1}{\delta} + \log\frac{ne^2}{d(f;\Delta)}\Big)$ is decreasing in Δ. Therefore,

$$\inf_{\Delta \geq \delta n^{-1/2} t^{\frac{1}{\alpha}+\frac{1}{2}}, \Delta \in \Delta_f} \varepsilon_n(d(f;\Delta);\delta;\Delta) \leq \varepsilon_n(d(f;\Delta_0);\delta;\Delta_0) \leq \frac{d(f;\bar{\Delta})}{n}\Big(\log\frac{1}{\delta} + \log\frac{ne^2}{d(f;\bar{\Delta})}\Big) + \frac{t}{n} \leq$$

$$\le \varepsilon_n(d(f;\bar\Delta);\delta;\bar\Delta) + \frac{t}{n} \le \varepsilon_n(f;\delta) + \frac{t}{n},$$

which proves (3.21).

Step 5. To complete the proof of the theorem, define the following event

$$E := \Big\{\exists f \in \mathcal{F}\ \exists \delta \in (0,1) : P_n\{f \le \delta\} \le \varepsilon_n(f;\delta) \text{ and } P\{f \le \frac{\delta}{4}\} \ge A\big(\varepsilon_n(f;\frac{\delta}{2}) + \frac{t}{n}\big)\Big\}.$$

Obviously, $E = E_1 \bigcup E_2$, where

$$E_1 := \Big\{\exists \delta \in (0,1)\ \exists f \in \mathcal{F}_1^\delta\ : P_n\{f \le \delta\} \le \varepsilon_n(f;\delta) \text{ and } P\{f \le \frac{\delta}{4}\} \ge A\big(\varepsilon_n(f;\frac{\delta}{2}) + \frac{t}{n}\big)\Big\},$$

$$E_2 := \Big\{\exists \delta \in (0,1)\ \exists f \in \mathcal{F}_2^\delta\ : P_n\{f \le \delta\} \le \varepsilon_n(f;\delta) \text{ and } P\{f \le \frac{\delta}{4}\} \ge A\big(\varepsilon_n(f;\frac{\delta}{2}) + \frac{t}{n}\big)\Big\}.$$

We set $\delta_j := 2^{-j}$, $j \ge 0$ and

$$\bar E_2 := \Big\{\exists j \ge 0\ \exists f \in \mathcal{F}_2^{\delta_j}\ : P_n\{f \le \delta_j\} \le \varepsilon_n(f;\delta_j) \text{ and } P\{f \le \frac{\delta_j}{2}\} \ge A\big(\varepsilon_n(f;\delta_j) + \frac{t}{n}\big)\Big\}.$$

It is easily seen that $E_2 \subset \bar E_2$. It follows from (3.20) that

$$\mathbb{P}(E_2) \le \mathbb{P}(\bar E_2) \le \sum_{j=0}^\infty \mathbb{P}\Big\{\exists f \in \mathcal{F}_2^{\delta_j} :\ P_n\{f \le \delta_j\} \le \varepsilon_n(f;\delta_j)$$
$$\text{and } P\{f \le \frac{\delta_j}{2}\} \ge A\big(\varepsilon_n(f;\delta_j) + \frac{t}{n}\big)\Big\} \le \sum_{j=0}^\infty B\delta_j^{1/8} e^{-t/4} \le B' e^{-t/4}.$$

If $f = \sum \lambda_i h_i \in \mathcal{F}_1^\delta$ for some δ then

$$\varepsilon_n(f,\delta) = \Big(\frac{\Delta(f)}{\delta}\Big)^{\frac{2\alpha}{2+\alpha}} n^{-\frac{2}{2+\alpha}} \bigvee \frac{2\log n}{n}.$$

where $\Delta(f) := \sum |\lambda_i|$. Therefore with some constant A'

$$E_1 \subseteq E_1' := \Big\{\exists \delta \in (0,1)\ \exists f \in \mathcal{F}\ P_n\{f \le \delta\} \le \Big(\frac{2\Delta(f)}{\delta}\Big)^{\frac{2\alpha}{2+\alpha}} n^{-\frac{2}{2+\alpha}} \bigvee \frac{2\log n}{n}$$
$$\text{and } P\{f \le \frac{\delta}{4}\} \ge A'\Big(\Big(\frac{\Delta(f)}{\delta}\Big)^{\frac{2\alpha}{2+\alpha}} n^{-\frac{2}{2+\alpha}} \bigvee \frac{2\log n}{n} + \frac{t}{n}\Big)\Big\}.$$

Let us first consider the case when the class $\mathcal{H}$ is uniformly bounded (say, by constant 1). One can observe that $\mathcal{F}' = \{f/\Delta(f) : f \in \mathcal{F}\} \subset \{f \in \mathrm{conv}(\mathcal{H}) : \Delta(f) = 1\}$. For any function f and any $\delta \ge \Delta(f)$, $P(f \le \delta) = 1$, which means that on the event E_1' one has to take into account only values of $\delta \le \Delta(f)$, or, equivalently, $\delta/\Delta(f) \le 1$. Therefore, a simple rescaling $\delta' = \delta/\Delta(f) < 1$ shows that

$$E_1' = \Big\{\exists \delta \in (0,1)\ \exists f \in \mathcal{F}'\ P_n\{f \le \delta\} \le \Big(\frac{2}{\delta}\Big)^{\frac{2\alpha}{2+\alpha}} n^{-\frac{2}{2+\alpha}} \bigvee \frac{2\log n}{n} \text{ and}$$

$$P\{f \le \frac{\delta}{4}\} \ge A\Big(\Big(\frac{1}{\delta}\Big)^{\frac{2\alpha}{2+\alpha}} n^{-\frac{2}{2+\alpha}} \bigvee \frac{2\log n}{n} + \frac{t}{n}\Big)\Big\}.$$

As to the second condition on $\mathcal{F}$, in this case $\Delta(f) = 1$ for any f by definition, and the above equivalent representation of the event E_1' holds automatically.

Let $\delta_j = 2^{-j}$, $j \ge 0$. Theorem 2 (see also Example 1) and a bound similar to (3.17) immediately imply that for some A and B

$$\mathbb{P}\Big\{\exists j\ \exists f \in \mathcal{F}'\ P_n\{f \le \delta_j\} \le \Big(\frac{1}{\delta_j}\Big)^{\frac{2\alpha}{2+\alpha}} n^{-\frac{2}{2+\alpha}} \bigvee \frac{2\log n}{n} \text{ and}$$

$$P\{f \le \frac{\delta_j}{2}\} \ge A\Big(\Big(\frac{1}{\delta_j}\Big)^{\frac{2\alpha}{2+\alpha}} n^{-\frac{2}{2+\alpha}} \bigvee \frac{2\log n}{n} + \frac{t}{n}\Big)\Big\} \le$$

$$\le \sum_{j\ge 0} B\exp\Big\{-\frac{1}{4}\Big(\frac{\sqrt{n}}{\delta_j}\Big)^{\frac{2\alpha}{2+\alpha}}\Big\} e^{-t/2} \le B' e^{-t/2}.$$

The same argument as before yields $\mathbb{P}(E_1') \le Be^{-t/2}$. Therefore, combining previuos bounds, we get $\mathbb{P}(E) \le Be^{-t/4}$, which completes the proof of the theorem.

□

4 Some experiments with learning algorithms

In this section we present some results of the experiments we conducted to test the ability of the new bounds to predict the value of the generalization error of combined classifiers. Unfortunately, the constants in the bounds of Section 2 are not known. More precisely, using the results of the recent work of Massart (1998) (Massart, 1998) one can calculate the constants involved in the bounds, but their current values are rather large and are way to far from being optimal. However, many important learning algorithms (such as boosting and bagging) that combine simple classifiers are iterative in nature and it's important to see whether the bounds allow one to predict the shape of the learning curves (the dependence of the generalization error on the number of iterations) correctly. To this end, we just ignore the constants and use in the experiments the quantities $(n^{1-\gamma/2}\hat{\delta}_n(\gamma; f)^\gamma)^{-1}$ (see Example 1) and $\varepsilon_n(f; \hat{\delta}_n(f))$ (see Theorem 3[2]) instead of the upper bounds we proved. We will refer to these quantities as the γ-bound and the Δ-bound, respectively. Incidentally, these quantities did provide upper bounds on the generalization error (or on the test error) in most of our experiments. This suggests that the values of the constants involved in the bounds of Section 2 might actually be moderate (at least in the case when the bounds are applied to several well known learning algorithms).

[2]Actually, the quantity $\varepsilon_n(f; \hat{\delta}_n(f)/2)$ is involved in this bound; but it's easy to see that it is within a constant from $\varepsilon_n(f; \hat{\delta}_n(f))$

4.1 Bagging and Boosting

We begin by describing the experiments with two of the most popular techniques of combining the classifiers, namely bagging (Breiman, 1996) and the Adaboost algorithm (Freund and Schapire, 1997). In both of these methods, there is an access to a learning algorithm called a *base learner*. The base learner is given a training sample $(X_i, Y_i),\ i = 1, \dots, n$ and it returns a classifier h from a base class $\mathcal{H}$ that "approximately minimizes" the empirical error $P_n\{yh(x) \le 0\}$ (or properly weighted empirical error).

In the case of bagging, the base learner receives at each iteration $t,\ t = 1, \dots, T$ an independent bootstrap sample $(\hat{X}_i^{(t)}, \hat{Y}_i^{(t)}),\ i = 1, \dots, n$ and returns a classifier $h_t \in \mathcal{H}$. The output of bagging is the combined classifier $f := T^{-1} \sum_{t=1}^{T} h_t$ (in other words, bagging makes a decision by majority vote).

In the case of Adaboost, the algorithm assigns at the beginning equal weights $D_1(i) = n^{-1},\ i = 1, \dots, n$ to all the training examples and then updates the weights iteratively. Namely, at t-th iteration $(t = 1, \dots, T)$ the algorithm calls the base learner that attempts to minimize approximately the weighted training error

$$\epsilon_t(h) := \sum_{i: h(X_i) \neq Y_i} D_t(i),\ h \in \mathcal{H}.$$

The base learner returns a classifier $h_t \in \mathcal{H}$ and its weighted training error $\hat{\epsilon}_t := \epsilon_t(h_t)$. The weights are then updated according to the formula

$$D_{t+1}(i) := \frac{D_t(i)}{Z_t}\big(1 + (\beta_t - 1) I_{\{h(X_i) = Y_i\}}\big),$$

where $\beta_t := \frac{\hat{\epsilon}_t}{1 - \hat{\epsilon}_t}$ and Z_t is the normalizing factor such that $\sum_{i=1}^{t} D_{t+1}(i) = 1$. After T iterations, Adaboost outputs a combined classifier

$$f := \Big(\sum_{i=1}^{T} \log \frac{1}{\beta_t}\Big)^{-1} \sum_{i=1}^{T} \log \frac{1}{\beta_t} h_t.$$

In all the experiments, we used the set of indicator functions[3] of axis oriented hyperplanes (also known as decision stumps) as base classifiers. That is, $S := \mathbb{R}^d$ and

$$\mathcal{H} = \big\{I_{\{\mathbf{x} \in \mathbb{R}^d : x_i \le c\}}, c \in \mathbb{R},\ i = 1, \dots, d\big\} \cup \big\{I_{\{\mathbf{x} \in \mathbb{R}^d : x_i \ge c\}}, c \in \mathbb{R},\ i = 1, \dots, d\big\},$$

where $\mathbf{x} = (x_1, \dots, x_d) \in \mathbb{R}^d$.

4.2 Experiments with real and simulated data

We first describe the experiments with a "toy" problem which is simple enough to allow one to compute exactly the generalization error and other quantities such as the γ-margins. Namely, we consider a one dimensional classification problem in which $S = [0, 1]$ and, given

[3] Actually, these functions are rescaled so that they take values in $\{-1, 1\}$

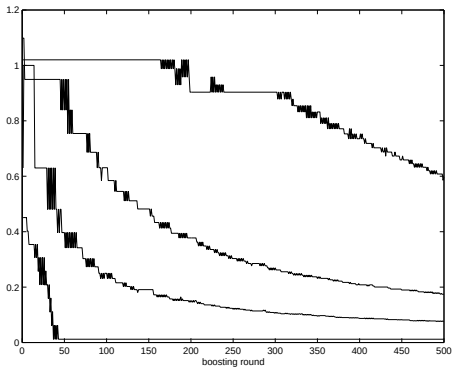


Figure 1: Comparison of the generalization error (thicker line) with $(n^{1-\gamma/2}\hat{\delta}_n(\gamma; f)^\gamma)^{-1}$ for $\gamma = 1, 0.8$ and $2/3$ (thinner lines, top to bottom).

a set (or a concept, using the terminology of computer learning) $C_0 \subset S$ which is a finite union of disjoint intervals, the label y is assigned to a point $x \in S$ according to the rule $y = f_0(x)$, where f_0 is equal to $+1$ on C_0 and to -1 on $S \setminus C_0$. We refer to this problem as the *intervals problem* (see also (Kearns et al., 1997)) Note that for the class of decision stumps we have in this case $V(\mathcal{H}) = 2$ (since $\mathcal{H} = \{I_{[0,b]} : b \in [0,1]\} \cup \{I_{[b,1]} : b \in [0,1]\}$), and according to the results above the values of γ in $[2/3, 1)$ provide valid bounds on the generalization error in terms of γ-margins. In our experiments, the set C_0 was formed by 20 equally spaced intervals and we generated a uniformly distributed on $[0,1]$ sample of size 1000. We ran Adaboost for 500 rounds (bagging does not work well for this problem), and computed at each round the generalization error of the combined classifier and the quantity $(n^{1-\gamma/2}\hat{\delta}_n(\gamma; f)^\gamma)^{-1}$ for different values of γ.

In figure 1 we plot the generalization error and the bounds for $\gamma = 1, 0.8$ and $2/3$ against the iteration of Adaboost. As expected, for $\gamma = 1$ (which corresponds roughly to the bounds in (Schapire et al., 1998)) the bound is very loose, and as γ decreases, the bound gets closer to the generalization error. In figure 2 we show that by reducing further the value of γ we get a curve that is even closer to the actual generalization error (although, for $\gamma = 0.2$, it does not provide an upper bound for some of the rounds of Adaboost). This seems to support the conjecture that Adaboost actually generates combined classifiers that belong to a subset of the convex hull of $\mathcal{H}$ with a smaller random entropy than of the whole convex hull. In figure 3 we plot the ratio $\hat{\delta_n}(\gamma; f)/\delta_n(\gamma; f)$ for $\gamma = 0.4, 2/3$ and 0.8 against the boosting iteration. We can see that the ratio is close to one in different examples (for a small number of iterations of Adaboost in the first example, the ratio is actually close to 0) indicating that the value of the constant $\bar{A}$ in the bound (2.5) might be close to one (at least, this seems to be true in the case of classifiers produced by Adaboost for large sample sizes).

In figure 4 we compare the γ-bound and the Δ-bound obtained for this problem for sample size of 1000. We can see that the Δ-bound has two regimes. In the first regime, the effect of the Δ-dimension is dominant, and the bound tracks almost exactly the generalization error, giving a definite improvement over the γ-bound. In the second regime, the bound starts increasing until it reaches the curve of the γ-bound. This behavior can be explained

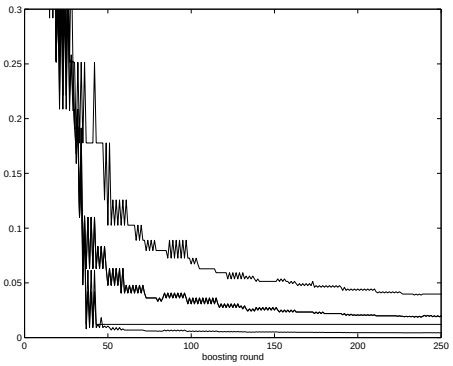


Figure 2: Comparison of the generalization error (thicker line) with $(n^{1-\gamma/2}\hat{\delta}_n(\gamma; f)^\gamma)^{-1}$ for $\gamma = 0.5, 0.4$ and 0.2 (thinner lines, top to bottom).

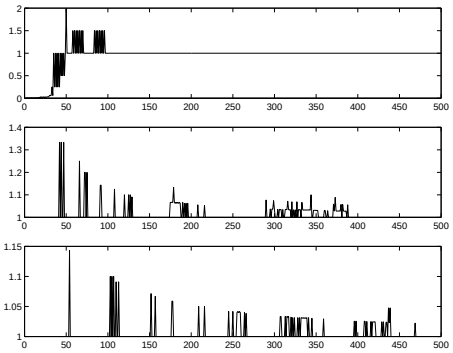


Figure 3: Ratio $\hat{\delta_n}(\gamma; f)/\delta_n(\gamma; f)$ versus boosting round for $\gamma = 0.4, 2/3, 0.8$ (top to bottom)

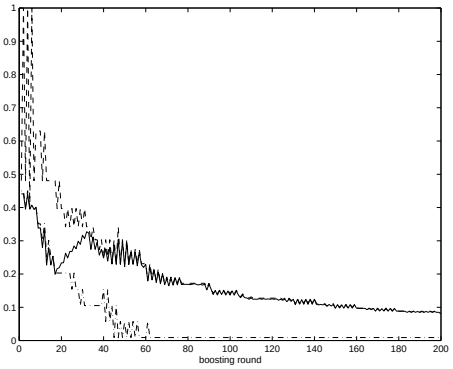


Figure 4: Test error and bounds vs. number of classifiers for the intervals problem for samples size of 1000. Test error (dot-dashed lines), γ-margin bound with $\gamma = 2/3$ (dashed lines), and Δ-bound (solid lines)

by examining the expression being minimized in the computation of the bound:

$$\underbrace{\frac{d(f;\Delta)}{n}\left(\log\frac{1}{\delta}+\log\frac{ne^2}{d(f;\Delta)}\right)}_{I}+\underbrace{\left(\frac{\Delta}{\delta}\right)^{\frac{2\alpha}{\alpha+2}}n^{-\frac{2}{\alpha+2}}}_{II} \tag{4.1}$$

It is easy to see that this expresion will be close to the γ-bound when the second term is dominant, and in fact, becomes the γ-bound when $\Delta = 1$ (which, apparently, is the case in our experiments when the number of classifiers in the convex combination becomes large).

We also computed the bounds for more complex simulated data sets as well as for real data sets in which the same type of behavior was observed. We show the results for the so called Twonorm Data Set and the King Rook vs. King Pawn Data Set (figure 5), which are well known examples in computer learning literature. The Twonorm Data Set (taken from (Breiman, 1998)) is a simulated 20 dimensional data set in which positive and negative training examples are drawn from the multivariate normal distributions with unit covariance matrix centered at $(2/\sqrt{20},\ldots,2/\sqrt{20})$ and $(-2/\sqrt{20},\ldots,-2/\sqrt{20})$, respectively. The King Rook vs. King Pawn Data Set is a real data set from the UCI Irvine repository (Blake and Merz, 1998)). It is a 36 dimensional data set with the sample size 3196.

As before, we used the decision stumps as base classifiers. An upper bound on $V(\mathcal{H})$ for the class $\mathcal{H}$ of decision stumps in $\mathbb{R}^d$ is given by the smallest n such that $2^{n-1} \geq (n-1)d+1$. We computed the Δ-bound and the γ-bounds for $\gamma = 1$ and for the smallest γ allowed in Example 1. For the Twonorm Data Set, we estimated the generalization error by computing the empirical error on an indepedently generated set of 20000 observations. For the King Rook vs. King Pawn Data Set, we randomly selected 90% of the data for training and used the remaining 10% to compute the test error. The experiments were averaged over 10 repetitions.

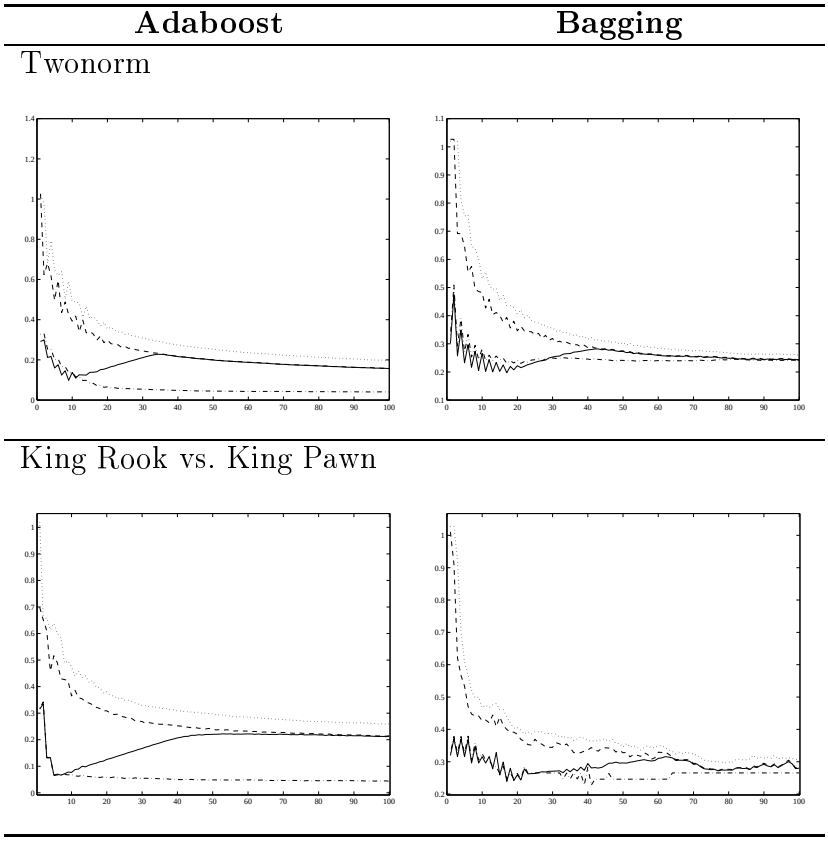


Figure 5: Test error and bounds vs. number of classifiers. Test error (dot-dashed lines), γ-margin bound with $\gamma = 1$ (dotted lines), and γ (dashed lines), and bbound (solid lines)

4.3 Weighting and normalization

It is apparent from the previous experiments that the Δ-bound explains well the behavior of the generalization error for a small number of classifiers in a convex combination, but for larger numbers of classifiers it becomes close to γ-bound. Partially, it might be related to the way the Δ-dimension was defined. In fact, the classifiers h_t output by the base learner at different iterations of Adaboost (or other voting method of combining classifiers) can be close to each other on the training examples (say, with respect to the distance $d_{P_n,2}$). Because of this, the Δ-dimension may very well overestimate the dimensionality of the combined classifier and more subtle definitions of dimension that take into account such empirical closeness of different functions in the convex combination are needed. The analysis of the proof of Theorem 3 shows that the extension of our bounds to these more subtle dimensions poses rather hard problems.

It might be also the case that the two terms in the expression (4.1) should be weighted in a certain way in order to obtain a better bound. The theoretical analysis of this problem is related to determining sharp values of the constants involved in the proof of Theorem 3 (which, in turn, is related to the problem of optimizing the constants in Talagrand's concentration and deviation inequalities for empirical processes that were used in the proof). We performed some experiments in order to study how such weighting influence the bound. More precisely, given $\zeta \in [0,1]$ and $K > 0$, we defined

$$\varepsilon_{n,\zeta,K}(f;\delta) := K \inf_{\Delta\in[0,1]}\Big[\frac{\zeta d(f;\Delta)}{n}\Big(\log\frac{1}{\delta}+\log\frac{ne^2}{d(f;\Delta)}\Big)+(1-\zeta)\Big(\frac{\Delta}{\delta}\Big)^{\frac{2\alpha}{\alpha+2}}n^{-\frac{2}{\alpha+2}}\Big]$$

We also looked at a possibility of "normalizing" the value of the Δ-dimension in the bound with respect to the total number of classifiers T :

$$\tilde{\varepsilon}_{n,\zeta,K}(f;\hat{\delta}_n(f)) := K \inf_{\Delta\in[0,1]}\Big[\frac{\zeta d(f;\Delta)/T}{n}\Big(\log\frac{1}{\hat{\delta}_n(f)}+\log\frac{ne^2}{d(f;\Delta)}\Big)+(1-\zeta)\Big(\frac{\Delta}{\hat{\delta}_n(f)}\Big)^{\frac{2\alpha}{\alpha+2}}n^{-\frac{2}{\alpha+2}}\Big].$$

We computed the bounds when weighting is used and when both weighting and normalization are used. We ran experiments for both simulated and real data sets in which we computed weighted and normalized bounds for values of $\zeta = 0.1, 0.2, \ldots 0.9$. We show results for $\zeta = 0.1, 0.4$ and 0.9 in figure 6.

We found that weighting with a value of $\zeta = 0.1$ gives for most of the data sets a curve that resembles rather closely the test error curve, and does not present two different regimes as before. When ζ increases (for example, when it becomes 0.4) the two-regime behavior becomes more noticeable, although for ζ close to one the curves exhibit only a small overshoot after which their shape is similar to the shape of the test error curve.

When normalization is introduced, we get curves that are very close to the test error curve for most of the data sets (regardless of the value of parameter ζ). At the moment, we do not have any theoretical explanation of these results.

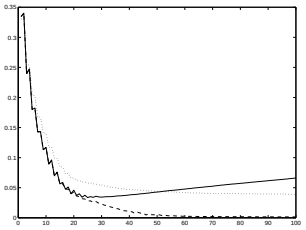
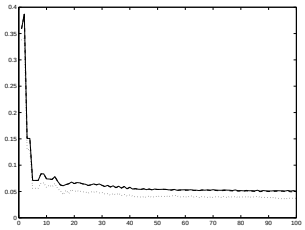
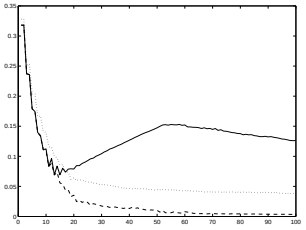
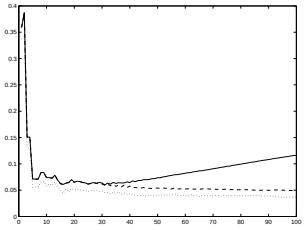
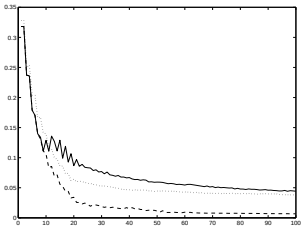
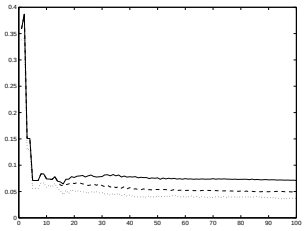
Twonorm	King Rook vs. King Pawn
$\zeta = 0.1$	
	
$\zeta = 0.4$	
	
$\zeta = 0.9$	
	

Figure 6: Bounds with weighting (solid line), weighting and normalization (dashed line) and test error (dotted line). In the bounds, $K = 1.14$.

4.4 Towards algorithms balancing the dimensionality and the margins

The connection between increasing the margins and reducing the generalization error has led to the development of several algorithms for designing and improving combined classifiers based on optimizing margin cost functions. The examples include DOOM (Mason et al., 2000), DOOM2 (Mason et al., 1999), DOOM-LP (Lozano and Koltchinskii, 2000), GeoLev (Duffy and Helmbold, 1999), and LP-Adaboost (Grove and Schuurmans, 1998). The results in this paper motivate the development of algorithms that take into account the approximate dimensions of combined classifiers along with their margins.

We discuss below the algorithm DOOM-LP, which was designed to optimize a piecewise linear cost function of the margins by solving a sequence of linear programs. Incidentally, this algorithm also tends to reduce the dimension of the combined classifier. To describe the algorithm, define $\varphi(u) := I_{(-\infty,0]}(u) + (1-u)I_{(0,1]}(u)$ and let $\varphi_\delta(u) := \varphi(u/\delta)$. Let $\mathcal{H}$ be a base class and $\mathcal{F} := \mathrm{conv}(\mathcal{H})$. It was proved in Koltchinskii and Panchenko (2000) that with probability at least $1 - 2\exp\{-2t^2\}$ the quantity

$$\inf_{\delta\in[0,1]}\left[P_n\varphi_\delta(yf(x)) + \frac{8}{\delta}\mathbb{E}\hat{R}_n(\mathcal{H}) + \left(\frac{\log\log_2(2\delta^{-1})}{n}\right)^{1/2}\right] + \frac{t}{\sqrt{n}}$$

is an upper bound on the generalization error $P\{yf(x)\le 0\}$ of *any* classifier $f\in\mathcal{F}$. Recall that $\hat{R}_n(\mathcal{H})$ is the Rademacher complexity of the class $\mathcal{H}$. If $\mathcal{H}$ is a VC-class, then $\mathbb{E}\hat{R}_n(\mathcal{H}) \le Cn^{-1/2}$ with a constant C depending on the VC-dimension of $\mathcal{H}$. The idea of the algorithm DOOM-LP is to minimize the above bound with respect to $f\in\mathcal{F}$ and $\delta\in[0,1]$ in order to find a classifier $\hat{f}$ with a reasonably small generalization error. More precisely, the algorithms receives a finite number of base classifiers $h_1,\dots,h_T$ along with their weights and attempts to redistribute the weights in order to minimize the bound.

For a fixed value of δ and fixed classifiers $h_1,\dots,h_T$, the minimization with respect to $f = \sum_{k=1}^T w_k h_k \in \mathcal{F}$ consists of finding the weights w_k, $\sum_{k=1}^T w_k = 1$, that minimize the following quantity:

$$P_n\varphi_\delta(yf(x)) = \frac{1}{n}\sum_{i=1}^n \varphi_\delta\left(Y_i\sum_{k=1}^T w_k h_k(X_i)\right). \tag{4.2}$$

For a given combined classifier $f = \sum_{k=1}^T w_k h_k \in \mathcal{F}$, define sets S_-, S_l, S_0 as follows:

$$S_- = \{i : Y_i f(X_i) \le 0\},\ S_l = \{i : 0 \le Y_i f(X_i) \le \delta\},\ S_0 = \{i : Y_i f(X_i) \ge \delta\}.$$

Finding the weight vector that "approximately minimizes" $P_n\varphi_\delta(yf(x))$ for a fixed current partition (S_-, S_l, S_0) can be easily posed as a linear programming problem. DOOM-LP searches for an approximate local minimum of $P_n\varphi_\delta(yf(x))$ by solving this linear program and moving to a neighboring partition by "flipping" the margins that fall in the intersection of two of the sets S_-, S_l, S_0 from the set they currently belong to another one in hope that with the constraints determined by the new partition the objective function can be reduced. The idea is similar in spirit to the sweeping hinge algorithm proposed by (Hush and Horne,

1998) Hush and Horn (1998). The algorithm converges when the value of the minimum in two neighboring partitions is the same (see algorithm 1). We use the following notations in the description of the algorithm: $b_k = -\sum_{i \in S_l} Y_i h_k(X_i)$ and $M_i = Y_i f(X_i)$, where $f = \sum_k w_k h_k$.

Algorithm 1 DOOM-LP

Require: Initial weight vector $\mathbf{w}$, margins $\{M_i\}_{i=1}^n$
  {Initialize the partition}
  $S_- = \{i : M_i \leq 0\}$
  $S_l = \{i : 0 \leq M_i \leq \delta\}$
  $S_0 = \{i : M_i \geq \delta\}$
  repeat
    $C_{min} = \sum_{k=1}^T b_k w_k$
    if $|S_l| \geq 1$ **then**
      {Compute optimal solution for a new partition}
      $\mathbf{w} = \text{LPSolve}(\mathbf{w}, S_-, S_l, S_0)$
      Compute new margins $\{M_i\}_{i=1}^n$
      {Update sets}
      $S_- = S_- \cup \{i : i \in S_l, M_i = 0\} - \{i : i \in S_-, M_i = 0\}$
      $S_l = S_l \cup \{i : i \in S_-, M_i = 0\} \cup \{i : i \in S_0, M_i = \delta\}$
        $-\{i : i \in S_l, M_i = 0 \text{ or } M_i = \delta\}$
      $S_0 = S_0 \cup \{i : i \in S_l, M_i = \delta\} - \{i : i \in S_0, M_i = \delta\}$
      $C = \sum_{k=1}^T b_k w_k$
    else
      Terminate and return current $\mathbf{w}$
    end if
  until $C \geq C_{min}$

If written in a standard form, the linear program solved by DOOM-LP at each iteration involves $T + n + |S_l| + 1$ variables (T weights plus slack and surplus variables) and $n + |S_l| + 1$ equality constraints. It follows from the basic results on linear programming that if there is an optimal feasible solution and the constraint matrix is full rank, then there exists an optimal feasible solution with at most $n + |S_l| + 1$ non zero variables. Furthermore, if the simplex method is used to solve the linear program, a solution of this type is allways found. We have observed in experiments that many of the variables that are set to zero in the solution are weights and that DOOM-LP tends to reduce the Δ-dimension of the classifier.

We have used DOOM-LP to improve the generalization error of combined classifiers produced by Adaboost by redistributing the weights of the base classifiers in a convex combination. An example of dimensionality reduction by DOOM-LP is illustrated in figure 7.

It might be interesting to design new algorithms with explicit penalization for high dimensionality in the optimization procedure. For instance, assuming that the initial weights $w_t^{(0)}, t = 1, \ldots T$ are arranged in decreasing order, one can add to the target function of linear program a term $\sum_{t=1}^T a_t w_t$, where $\{a_t, t \geq 1\}$ is an increasing sequence of positive numbers. One can also consider entropy type penalties of the form $\sum_{t=1}^T w_t \log \frac{1}{w_t}$ (in this case, of course, the optimization is not a linear programming problem any longer).

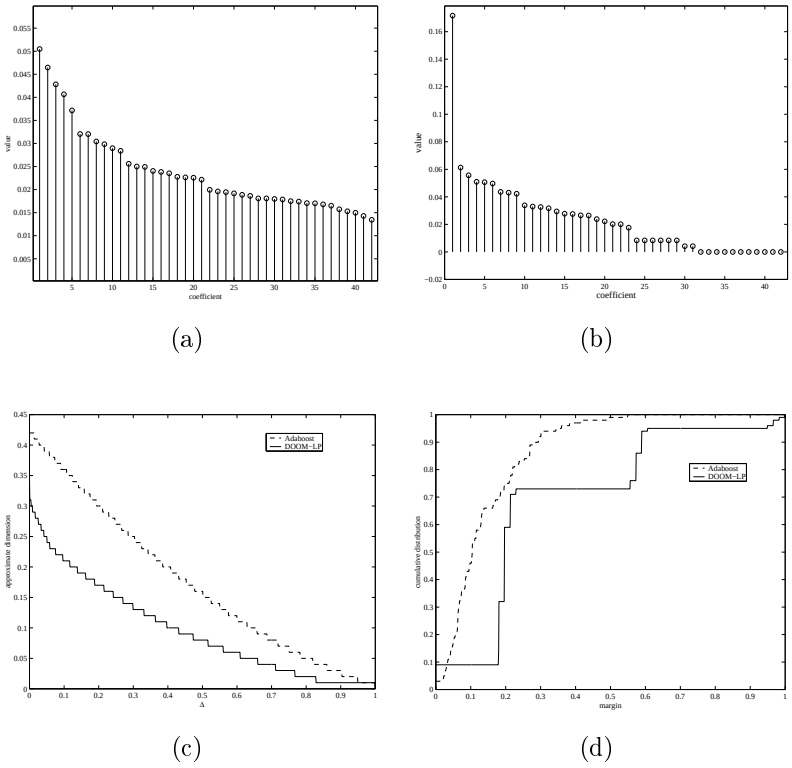


Figure 7: Results of running DOOM-LP on the classifier produced by Adaboost for the King Rook Vs. King Pawn data set. (a) Adaboost sorted coefficients, (b) DOOM-LP sorted coefficients, (c) Approximate Δ-dimensions, (d) Cumulative margin distributions.

References

Anthony, M. and Bartlett, P. (1999). *Neural network learning:theoretical foundations*. Cambridge University Press.

Bartlett, P. (1998). The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Transactions on information theory*, 44:525–536.

Bartlett, P., Boucheron, S., and Lugosi, G. (2000). Model selection and error estimation. Preprint.

Blake, C. and Merz, C. (1998). UCI repository of machine learning databases. URL: http://www.ics.uci.edu/∼mlearn/MLRepository.html.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 26:123–140.

Breiman, L. (1998). Arcing classifiers. *The Annals of Statistics*, 26(3):801–849.

Cortes, C. and Vapnik, V. (1995). Support vector networks. *Machine Learning*, 24:273–297.

Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New York.

Duffy, N. and Helmbold, D. (1999). A geometric approach to leveraging weak learners. In *Eurocolt99*.

Freund, Y. and Schapire, R. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139.

Grove, A. and Schuurmans, D. (1998). Boosting in the limit: maximizing the margin of learned ensembles. In *Proceedings of the fifteenth national conference on Artificial intelligence*.

Hush, D. and Horne, B. (1998). Efficient algorithms for function approximation with piecewise linear sigmoids. *IEEE Transactions on Neural Networks*, 9(6):1129–1141.

Kearns, M., Mansour, Y., Ng, A., and Ron, D. (1997). An experimental and theoretical comparison of model selection methods. *Machine Learning*, 27(1).

Koltchinskii, V. (1999). Rademacher penalties and structural risk minimization. *Submitted to IEEE Transactions on Information Theory*.

Koltchinskii, V. and Panchenko, D. (1999). Empirical margin distribution and bounding the generalization error of combined classifiers. Preprint.

Koltchinskii, V. and Panchenko, D. (2000). Rademacher processes and bounding the risk of function learning. In E. Gin´ e, D. M. and Wellner, J., editors, *High Dimensional Probability II*, pages 00–00, Boston. Birkhäuser.

Koltchinskii, V., Panchenko, D., and Lozano, F. (2000). Bounding the generalization error of neural networks and combined classifiers. In *Advances in Neural Information Processing Systems 13*. to appear.

Lozano, F. and Koltchinskii, V. (2000). Direct optimization of simple cost functions of the margin. Preprint.

Mason, L., Bartlett, P., and Baxter, J. (2000). Improved generalization through explicit optimization of margins. *Machine Learning*, 38(3):243–255.

Mason, L., Baxter, J., Bartlett, P., and Frean, M. (1999). *Advances in large margin classifiers*, chapter Functional gradient techniques for combining hypotheses. MIT Press.

Massart, P. (1998). About the constants in Talagrand's concentration inequalities for empirical processes. Preprint, Université Paris-Sud.

Pisier, G. (1981). Remarques sur un résultat non publié de b.maurey. séminaire d'analyse fonctionelle. In *Séminaire d'analyse Fonctionelle, 1980-1981, Exposé No. 5.*

Schapire, R., Freund, Y., Bartlett, P., and Lee, W. (1998). Boosting the margin : A new explanation for the efectiveness of voting methods. *Annals of Statistics*, 26(5):1651–1687.

Talagrand, M. (1996a). New concentration inequalities in product spaces. *Invent. Math.*, 126:505–563.

Talagrand, M. (1996b). A new look at independence. *Annals of Probability*, 24:1–34.

van der Vaart, A. W. and Wellner, J. (1996). *Weak Convergence of Empirical Processes With Applications to Statistics.* Springer Series in Statistics. Springer.

Vapnik, V. (1998). *Statistical Learning Theory.* John Wiley and Sons, Inc.